Rapport de stage

Analyse statistique de trajectoires de véhicules en virage

Pierre-Emmanuel Caute
Stagiaire au Laboratoire sur les Interactions Véhicules-Infrastructure-Conducteurs

27 septembre 2010

Remerciements

Je tiens tout d'abord à remercier Guillaume Saint Pierre qui a accepté de m'encadrer. Il a su garder sa confiance en mon travail, tout en se rendant disponible à chaque fois que j'avais besoin d'aide.

J'adresse mes remerciements respectueux à Jacques Ehrlich pour m'avoir accueilli au LIVIC dont il est le directeur.

Je remercie toute l'équipe du LIVIC et particulièrement Cindie pour tous ses conseils.

Enfin un grand merci à mes collègues stagiaires Mai Linh, Soraya, Alix, Antoine et Vincent pour avoir contribué à créer une ambiance de travail agréable.

Table des matières

Liste des figures

Liste des tableaux

Résumé

Dans un contexte où 40% des accidents mortels en France sur les routes de campagne se produisent en courbe, les innovations de l'assistance à la conduite peuvent être motivées par une analyse préalable des trajectoires en virage.

Un "Observatoire de Trajectoires" a été développé par Yann Goyat pour mesurer de façon précise l'évolution de la vitesse et de la position des véhicules. Les données qu'il fournit ont conduit à des propositions de typologies des conduites. Ce rapport vise à évaluer la qualité de l'une de ces typologies.

Quatre méthodes de classification mettant en jeu des outils statistiques tels que l'analyse en composantes principales, la régression logistique, les classifications hiérarchiques ascendante et descendante et la méthode des K-moyennes ont été appliquées à un échantillon de trajectoires afin d'être comparées. Un indicateur de risque fondé sur la position des véhicules a été proposé.

Abstract

Within a framework where 40% of fatal accidents on country roads in France occur on bends, innovations in assisted driving can be supported by a preliminary analysis of turning trajectories.

A "trajectory observatory" was developed by Yann Goyat to accurately measure the changes of speed and vehicle position. The provided data has lead to propositions of driving typologies. This report aims to evaluate the quality of one of these typologies.

Four classification methods using statistical tools such as the analysis of principal components, logistic regression, hierarchical ascending and descending classification, k-means method have been applied to a sample of trajectories for comparison. A proposed risk indicator has been based on the position of the vehicle.

Chapitre 1

Présentation du LIVIC

Le LIVIC est une unité mixte INRETS/LCPC créée en mars 1999. Ses activités de recherche sont destinées à améliorer le fonctionnement des réseaux routiers par le développement de systèmes technologiques d'aides à la conduite permettant une meilleure coopération entre les conducteurs, les véhicules et l'infrastructure.

1.1 Les moyens humains

Le Laboratoire sur les Interactions Véhicules-Infrastructure-Conducteurs dispose actuellement d'un effectif variable de 20 à 30 personnes constitué de personnel administratif et de direction, de chercheurs, d'ingénieurs et techniciens, de contractuels, de doctorants et de stagiaires. Cet effectif est réparti en cinq équipes : administrative, perception, contrôle-commande, systèmes coopératifs, et validation expérimentale.

1.2 Des moyens matériels en évolution

1.2.1 Les bureaux

Ils se situent au bâtiment N°824 du site du GIAT. Ce bâtiment permet d'accueillir 30 agents sur deux étages. Il contient également

- un laboratoire d'électronique (LABELEC)
- un laboratoire de traitement d'image (LABTI)
- une salle de bibliothèque et de lecture
- deux salles de réunion
- une cuisine

1.2.2 L'atelier de transformation des véhicules

C'est un bâtiment de 300 m2 dont l'aménagement a été effectué en 2004. Il est divisé en aires distinctes permettant l'accomplissement de travaux de diverses disciplines.

- mécanique
- électricité, électronique et communications
- programmation des équipements embarqués
- salle de projection - banc d'essai

C'est un véritable atelier de recherche permettant des travaux logiciels et matériels sur les véhicules tout en étant connecté au réseau informatique comme dans son bureau.

1.2.3 Les pistes d'essais

Propriété du ministère de la Défense elles sont gérées par le groupe GIAT Industrie, elles font l'objet d'une convention d'utilisation. Elles se décomposent comme suit :

- une route rapide de 2 km environ,
- une route plus sinueuse de type nationale de 4 km (La Routière)
- une route en forêt de type chemin départemental de 2 km,
- un anneau de 300 m de diamètre,
- une aire de type parking de 1 ha
- une aire circulaire de 300 m de diamètre (dite la rotonde).

1.2.4 Quatre véhicules d'essais opérationnels

1.2.5 SIVIC : Un moyen virtuel de test

SiVIC, (Simulateur de Véhicules, d'Infrastructure et de Capteurs virtuels) est une plate forme de prototypage d'environnement routier et de capteurs virtuels pour les aides à la conduite. En l'absence de données réelles, SIVIC permet de produire des données provenant de simulations numériques. Il offre des solutions de substitution pour tester et valider les algorithmes de perception et de contrôle/commande qui entrent dans la conception des différents systèmes d'aide à la conduite développés.

1.3 La démarche scientifique au LIVIC

Compte tenu de moyens limités mis en place au LIVIC au regard de l'étendue du champ visé, le LIVIC se situe pour l'instant dans une recherche d'équilibre entre analyses conceptuelles, développements techniques et validations expérimentales.

5 thèmes ou lignes d'action sont mis en place :

- Analyses de concepts et évaluation a priori
- Développement de moyens de perception,
- Modélisation des véhicules et contrôle commande
- Validation expérimentale de la conduite automatisée ou partagée.
- Systèmes coopératifs

1.4 Les contrats du LIVIC

Le LIVIC inscrit une grande partie de ses activités dans le cadre de projets de recherche nationaux et européens.

- Contrats ARCOS (Action de Recherche pour une Conduite Sécurisée)
- Convention LAVIA
- Contrat européen PREVENT/SAFELANE
- Projet TRACKSS
- Projet LOVE

Chapitre 2

Présentation de l'expérience

Ce travail s'inscrit dans le projet PARTAGE (Contrôle partagé entre conducteur et assistance à la conduite automobile pour une trajectoire sécurisée), héritier du projet DIVAS, et plus globalement des projets scientifiques destinés à améliorer la sécurité routière.

2.1 Du Dialogue Infrastructure - Véhicules pour Améliorer la Sécurité routière (DIVAS) à l'analyse de trajectoires en virages

2.1.1 Le projet DIVAS

Le projet DIVAS est né de l'observation des chiffres de l'insécurité routière, et de leur possible réduction par le développement d'échanges d'information en temps réels entre infrastructure et véhicules. Son objectif est de bâtir une conception globale de système d'échanges infrastructure-véhicules efficace en termes de sécurité routière, et d'en préparer le déploiement en examinant toutes les conséquences, notamment en termes technologiques mais aussi sur les plans de la crédibilité et de l'acceptabilité. On rappelle qu'en France, 40% des accidents mortels sur les routes de campagne se produisent dans une courbe et que dans la plupart des cas, les principaux facteurs accidentogènes sont liés aux caractéristiques géométriques et de surfaces de la route. Il a donc semblé nécessaire d'observer les trajectoires prises dans certains virages à risques.

2.1.2 L'observations des trajectoires

2.1.2.1 L'Observatoire de Trajectoires (OdT)

L'observatoire de Trajectoire (OdT) est un système d'instrumentation composé de deux caméras, un télémètre à balayage laser et des cellules de détection. Une station de travail informatique permet de récupérer toutes les données enregistrées par ces capteurs. La photo 2.3 est un exemple de l'instrumentation réalisée sur le site du lieu dit de la Trocharderie, et montrant les caméras, installées en haut d'un mât à une hauteur maximale de 6m, le télémètre et l'armoire blindée comprenant le matériel informatique d'acquisition. Tout ceci est placé en bord de voie, vers le centre du virage, quand cela est possible. Les cellules de détection sont quant à elles placées en amont du virage, à une distance d'environ 70m de l'OdT. Elles ont pour rôles de détecter le passage des véhicules légers, et ainsi de déclencher

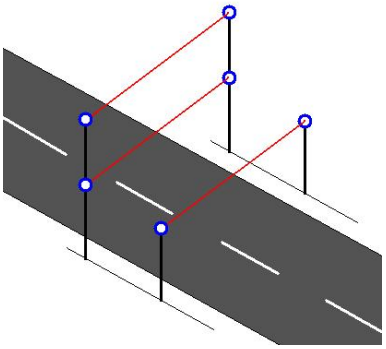


FIGURE 2.1 – *Disposition des cellules de détection*

l'enregistrement des caméras et du télémètre. Leur disposition se fait comme le montre la figure 2.1 Ceci permet de détecter le sens de passage du véhicule (2 cellules à même hauteur) et de détecter le passage d'un poids lourd (cellule haute). Le schéma de la figure 2.2 est une représentation des zones couvertes par les capteurs que sont les deux caméras et le télémètre. Cette disposition permet un suivi du parcours des véhicules sur une longueur efficace d'environ 100m.

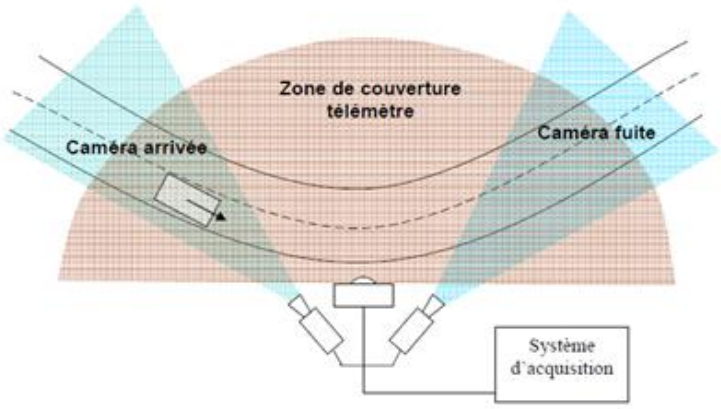


FIGURE 2.2 – *Disposition et zones de couverture des capteurs*

2.1.2.2 Le logiciel de traitement SAVe

Les fichiers enregistrés par l'OdT sont ensuite analysés en post traitement à l'aide du logiciel de suivi « SAVe Virage » . Il réalise un tracking des véhicules tout en fusionnant les données issues des enregistrements télémètres et celles issues des vidéos et ainsi, détermine pas à pas la valeur de nombreux paramètres tels que la vitesse, la position, le cap, l'angle de braquage, l'heure de passage, les dimensions du véhicules, etc ... Le logiciel étant encore en cours de développement pendant la durée des expérimentations, le traitement des enregistrements de l'OdT a été effectué avec différentes versions du logiciel. Les premières versions

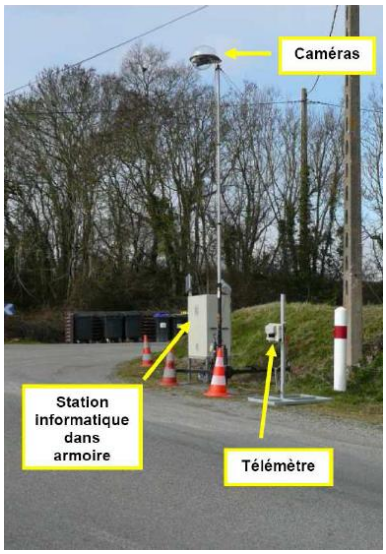


FIGURE 2.3 – *Photo de l'OdT*

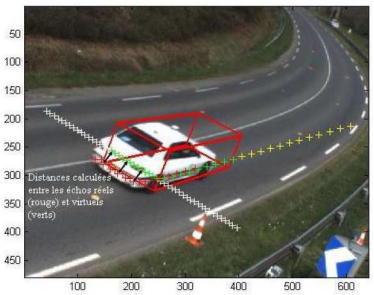


FIGURE 2.4 – *Mesure de la position et du cap d'un véhicule*

ne permettaient pas de fusionner les données des champs arrivée, fuite et télémètre pour un seul et même véhicule.

2.1.2.3 Présentation des données

L'échantillon est composé de 3007 trajectoires, décrites par les variables détaillées dans le tableau 2.1 en des observations dont le nombre varie.

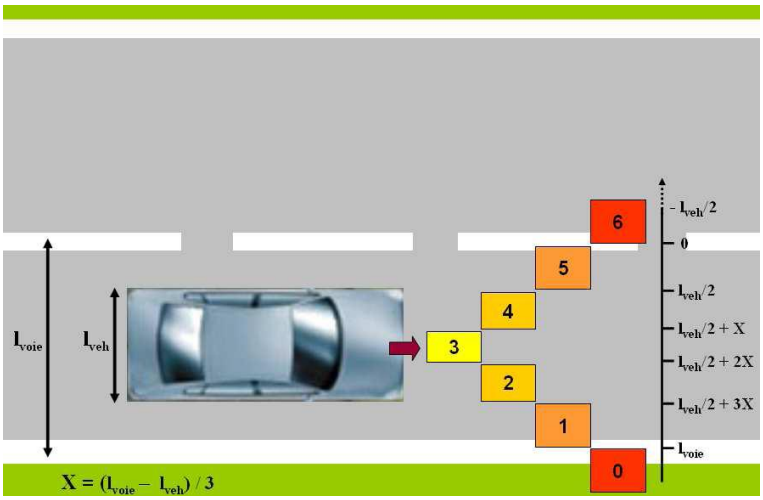


FIGURE 2.5 – *Explication de la variable Position, graphe tiré du livrable 441 du projet SARI (Goyat et Menant [2008])*

2.1.3 Analyse de trajectoires proposée dans DIVAS

Le projet DIVAS, comme d'autres projets, a donc utilisé les mesures de l'OdT pour l'analyse de trajectoires en virage.

Parmi les résultats de ce projet, nous observons sur le livrable 2A2 (Goyat et al. [2008b]) qu'une classification des trajectoires en 3 types de comportement est proposée. Ces 3 classes y sont ainsi présentées :

- 1er type : changement de trajectoire entre l'entrée et le milieu du virage. Les véhicules abordent le virage en position centrale et se décalent ensuite vers le marquage central (figure 2.6).

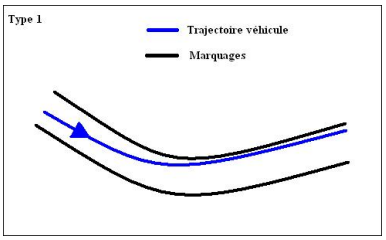


FIGURE 2.6 – *Type de conduite numéro 1*

- 2ème type : changement de trajectoire entre le milieu et la sortie du virage. Les véhicules abordent le virage proche du marquage central et se recentrent sur la voie après avoir passé le milieu du virage (figure 2.7).

Variable	Desciption
Numéro	Identifiant du trajet
Temps	Temps de la mesure (en 10^{-6} secondes)
Année	Année de la mesure
Mois	Mois de la mesure
Jour	Jour de la mesure
Heure	Heure de la mesure
Minute	Minute de la Mesure
Seconde	Seconde de la Mesure
Latitude	Latitude du véhicule, un offset lui est appliqué afin de rendre les valeurs plus maniables (en m)
Longitude	Longitude du véhicule, un offset lui est appliqué afin de rendre les valeurs plus maniables (en m)
Vitesse	Vitesse du véhicule, calculé par le SAVe (en km/h)
Position	Position du véhicule dépendant de l'écart au marquage central (de 6 : plus de la moitié du véhicule circule sur la voie opposée, à 0 : plus de la moitié du véhicule circule sur l'accotement) cf figure 2.5
Longueur	Longueur du véhicule (en m)
Largeur	Largeur du véhicule (en m)
Hauteur	Hauteur du véhicule (en m)
Cap	Cap pris par le véhicule (direction)
Angle Au Volant	Calculé à partir des Caps successifs
Ecart	Ecart au marquage central (en m)
Type	Type de véhicule (clair = 0, sombre = 1, camion = 2, moto = 3, indéfini = 4)
Intervalle de Temps	Différence entre deux Temps successifs
Sens	Sens de circulation
Champs	Champs de caméra (1 pour la caméra d'entrée, 2 pour télémètre,3 pour la camera de sortie, 4 pour la fusion caméra d'entrée - télémètre, et 5 pour la fusion caméra de sortie - télémètre)
Date Du Fichier	Date de création du fichier (en s)

TABLE 2.1 – *Les variables obtenues grâce à l'OdT*

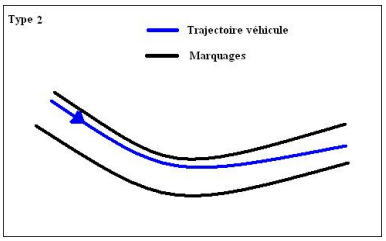


FIGURE 2.7 – *Type de conduite numéro 2*

– 3ème type : changement de trajectoire avant et après le centre du virage. Les véhicules abordent le virage en position centrale, puis se décalent vers la ligne médiane au milieu du virage avant de revenir à une position centrale sur la voie. Cette configuration est typique d'une volonté de la part de l'usager de vouloir limiter la décélération et augmenter son confort de conduite en adoptant une trajectoire la plus rectiligne possible, trouvant alors le point de corde au centre du virage (figure 2.8).

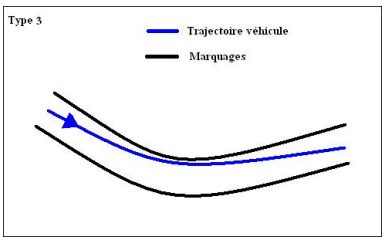


FIGURE 2.8 – *Type de conduite numéro 3*

2.2 L'expérience

L'objectif était d'associer une trajectoire à un niveau de risque. Pour ce faire, deux expériences étaient prévues dans le cadre de PARTAGE :

– 20 conducteurs devaient répondre à un questionnaire sensé traduire leur état d'esprit au moment de prendre le volant. Puis ils devaient parcourir un circuit plusieurs fois en allant de plus en plus vite, jusqu'à ce qu'ils souhaitent arrêter. Ce circuit présentait un virage équipé de l'OdT.
– L'Odt devait être installé dans ce même virage pendant plusieurs mois, et renvoyer les mesures de tous le véhicules qui l'empruntait

Pour des raisons techniques, les données n'ont pu être disponibles dans le cadre de ce stage. Nous avons donc travaillé avec un échantillon de trajectoires observées en juin 2009.

Les méthodes d'analyse retenues sont :

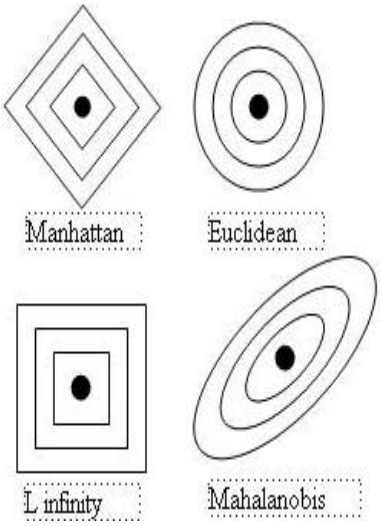
– la description des trajectoires à l'aide d'indicateurs, calculés sur chaque partie du virage (entrée, centre, sortie),
– la description des trajectoires comme une fonction du temps.

Chapitre 3

Outils mathématiques pour le traitement des trajectoires

Nous présenterons ici deux des outils mathématiques utilisés pour l'analyse des trajectoires : les distances, nécessaires au classement des trajectoires, et l'interpolation qui nous a permis de les mettre en forme.

3.1 Distances



3.1.1 Introduction

Afin de classer les trajectoires, il existe de nombreuses méthodes de classification. Nous nous intéressons ici aux algorithmes de classification non supervisée, tels que la classification hiérarchique ou la méthode des K-means, lesquels demandent un indice de dissimilarité. Ce document vise à présenter les indices pouvant remplir cette tâche, leur avantages et leurs inconvénients. Les trajectoires sont obtenues sous la forme de données longitudinales quan-

titatives, nous nous restreindrons donc à une présentation des dissimilarités entre données quantitatives.

3.1.2 Définition : Distance, Dissimilarité

Une métrique est une fonction binaire qui décrit la distance entre deux points d'un ensemble E. Cette distance est un application de $E \times E \to R^+$ telle que, :

$$\begin{cases} \text{Symétrie} : d(x,y) = d(y,x). \\ \text{Positivité} : d(i,j) \geq 0. \\ \text{Séparation} : d(x,y) = 0 \Leftrightarrow x = y. \\ \text{Inégalité triangulaire} : d(x,y) \leq d(x,z) + d(z,y). \end{cases}$$

On parle de dissimilarité quand on a seulement :

$$\begin{cases} d(x,y) = d(y,x), \\ d(x,y) \geq 0, \\ d(x,x) = 0. \end{cases}$$

3.1.3 Exemples de métrique

Etant donné un espace vectoriel normé $(X, ||.||)$, nous pouvons définir une métrique sur X par

$$d(x,y) := ||x - y||.$$

Une norme sur E est une application N sur E à valeurs réelles positives satisfaisant les hypothèses suivantes :

$$\begin{cases} \text{Séparation} : \forall x \in E, N(x) = 0 \Rightarrow x = 0_E. \\ \text{Homogénéité} : \forall (\lambda, x) \in K \times E, N(\lambda.x) = |\lambda| N(x). \\ \text{inégalité triangulaire} : \forall (x,y) \in E^2, N(x+y) \geq N(x) + N(y). \end{cases}$$

On trouve ainsi :

– La distance de Manhattan issue de la norme L^1 :
Aussi appelée distance « city-block » ou distance de Gower. elle est souvent réservée aux classifications hiérarchiques, il s'agit de la somme des valeurs absolues des distances. Elle ne majore donc pas la pondération des outliers. En revanche, les temps de calcul sont particulièrement longs.

$$Manh(x,y) = \sum_{i=1}^{n} |x_i - y_i|.$$

– La distance euclidienne issue de la norme L^2 :
C'est probablement le type de distance le plus couramment utilisé. Il s'agit simplement d'une distance géométrique dans un espace multidimensionnel.

$$Eucl(x,y) = \sqrt{(}\sum_{i=1}^{n}(x_i - y_i)^2).$$

Il est courant d'utiliser le carré de la distance euclidienne, plus rapide à calculer que la distance euclidienne pour des classifications par la méthode des K-means ou la méthode de Jarvis-Patrick, car leurs résultats ne sont pas affectés. Toutefois, pour les méthodes de classification hiérarchique, les objets les plus atypiques (éloignés) sont "sur-pondérés".

– La distance de Minkowski issue de la norme L^p :

$$Mink(x,y) = \sqrt[p]{(}\sum_{i=1}^{n}|x_i - y_i|^p).$$

Une généralisation de la distance de Minkowsky nous donne la distance de puissance :

$$Mink(x,y) = \sqrt[r]{(}\sum_{i=1}^{n}|x_i - y_i|^p).$$

Cette distance permet de jouer indépendamment sur les deux puissances présentes dans l'équation, pour trouver l'équilibre voulu entre l'importance du nombre d'éléments différents et l'importance de la différence elle-même.

– La distance de Chebyshev issue de la norme L^∞ :
La distance de Chebyshev mesure la distance maximale qui va exister entre deux points. Dans l'espace, un individu ayant une caractéristique extraordinaire sera plus isolé qu'un individu ayant plusieurs caractéristiques un peu particulières. Aux échecs, elle représente le nombre de coup qu'il faut à un roi pour aller jusqu'à une certaine case, comme le montre la figure 3.2 .

$$Cheb(x,y) = \sup|x_i - y_i|.$$

– La distance de Mahalanobis :
Elle a été établie par Prasanta Chandra Mahalanobis en 1936. Elle correspond à la distance euclidienne normalisée par la covariance. Elle ne depend pas de l'échelle des données. Elle est souvent utilisée pour obtenir des classes de forme ellipsoïdale (figure 3.3).

$$Maha(x,y) = \sqrt{((x-y)^T\Sigma^{-1}(x-y))}.$$

où Σ est la matrice de covariance des données.

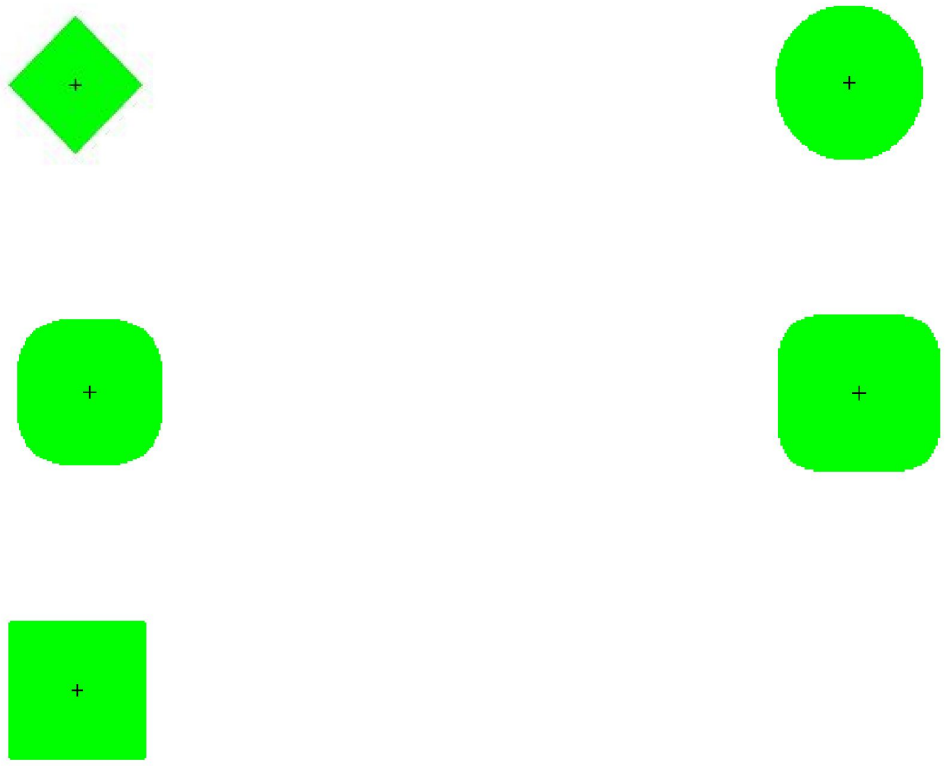


FIGURE 3.1 – *De gauche à droite et de haut en bas, distances de Manhattan, euclidienne, de Minkowski pour p=3, p=4, p=20.*

– La distance de Canberra :

$$Canb(x,y) = \sum_{i=1}^{n} \frac{|x_i - y_i|}{|x_i| + |y_i|}.$$

La distance de Canberra est très sensible aux changements proches de 0.

– La distance de Bray-Curtis :

$$Bray(x,y) = \frac{\sum_{i=1}^{n} |x_i - y_i|}{\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i}.$$

La distance de Bray Curtis, également appelée Distance de Sorensen, est une méthode de normalisation utilisée généralement en botanique, et autres sciences naturelles. Elle conçoit l'espace comme une grille, comme la distance de Manhattan. Lorsque toutes les coordonnées sont positives, elle prend sa valeur entre 0 et 1, le zero signifiant que les données sont identiques. Cette distance est indéfinie en 0.

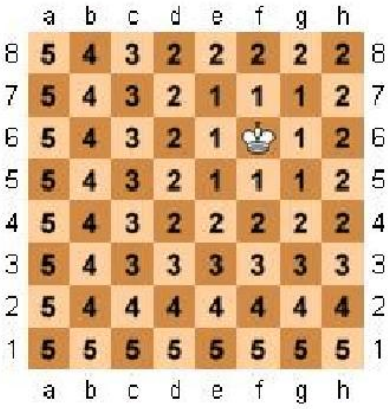


FIGURE 3.2 – *Distance de Chebyshev*

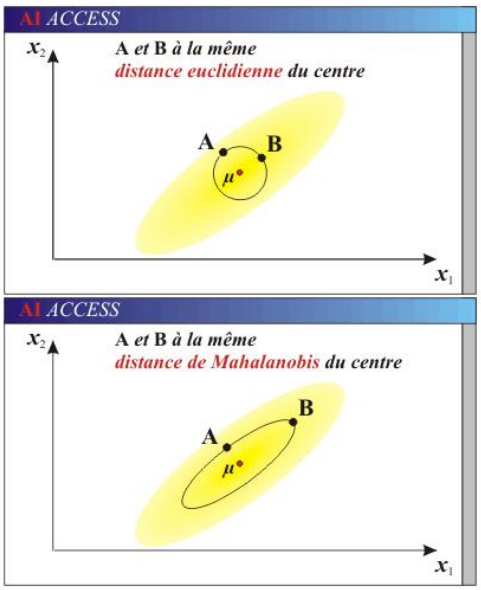


FIGURE 3.3 – *Distance de Mahalanobis*

– distance du Khi-deux :

$$khi2(x,y) = \sqrt{(\sum_{i=1}^{n} \frac{(x_i - y_i)^2}{|x_i + y_i|}}).$$

– distance des cordes carrées (squared cords distance) :

$$cord(x,y) = \sum_{i=1}^{n} (\sqrt{x}_i - \sqrt{y}_i)^2.$$

Toutes ces distances concernent des vecteurs de même taille, or les trajectoires dont nous disposons sont mesurées à des intervalles de temps réguliers. Comme les véhicules n'ont pas tous la même vitesse, la taille des observations varie. Il nous faut donc des solutions pour mesurer la distance entre deux vecteurs de tailles différentes.

3.1.4 Dynamic Time Warping

La distance dynamique par déformation temporelle ou dynamic time warping distance est utilisée en fouille de données de séries temporelles, notamment pour des problèmes de reconnaissance vocale. Cette distance compare chaque point avec les points voisins. Elle cherche un chemin de déformation à travers les valeurs de x et y qui minimise la somme des différences point à point. Un chemin est défini comme une suite de couples (i,j) pour lesquels x_i et y_j sont comparés. Soit $w = w_1, ..., w_s, ..., w_k = (i_s, j_s)_{s=1,...,k}$ un chemin de taille k. Nous notons W l'ensemble des chemins admissibles, c'est-à-dire les chemins répondant à une contrainte de continuité entre deux positions et une contrainte aux extrémités. La distance d_{dtw} s'exprime de la façon suivante :

$$d_{dtw}(x,y) = \min_{w \in W} \sum_{(i_s,j_s) \in w} d(x_{i_s}, y_{j_s}), W = \{w \in \{1,...,n\}^{2k}\},$$

$$\text{tel que } \begin{cases} (i_1, j_1) = (1,1) \text{ et } (i_k, j_k) = (n,n) \\ \forall s, i_s - i_{s-1} \geq 1 \text{ et } j_s - j_{s-1} \geq 1 \end{cases}$$

où d une métrique.

Elle est implémentée sous R dans le package "DTW" qui laisse l'utilisateur libre de choisir les contraintes et la distance.

3.1.4.1 Le choix de la distance

Les distances peuvent être celles proposées par le package "proxy" :
- Euclidienne,
- Manhattan,
- Canberra,
- Minkowski.

3.1.4.2 Le choix des contraintes

L'intérêt de la DTW est qu'elle permet de choisir quels points seront comparés pour la comparaisons de deux séries. Pour le choix de ces points, nous pouvons imposer que leurs indices ne soient pas trop éloignés, ou choisir la nature de cet éloignement. La fonction des indices de la première trajectoire selon les indices de la seconde appartient alors à une "fenêtre". De même nous pouvons choisir le poids de chaque choix d'indice : comme par exemple associer à une comparaison (i,j) un coefficient 1, tandis qu'à $(i-1,j)$ ou $(i,j-1)$, nous associons 2.

1. Les fenêtres :
   - La fonction de Sakoe Chiba implémente la bande de Sakoe Chiba, c'est-à-dire une fenêtre dont nous pouvons choisir la largeur symétrique autour de la première diagonale.
   - La fenêtre "slantedBand" (en bande inclinée) est une bande centrée autour du segment joignant l'élément [1,1] à au dernier élément de chaque trajectoire. On peut en choisir la largeur.
   - La fenêtre d'Itakura, est présentée, comme les autres fenêtres en figure 3.4.

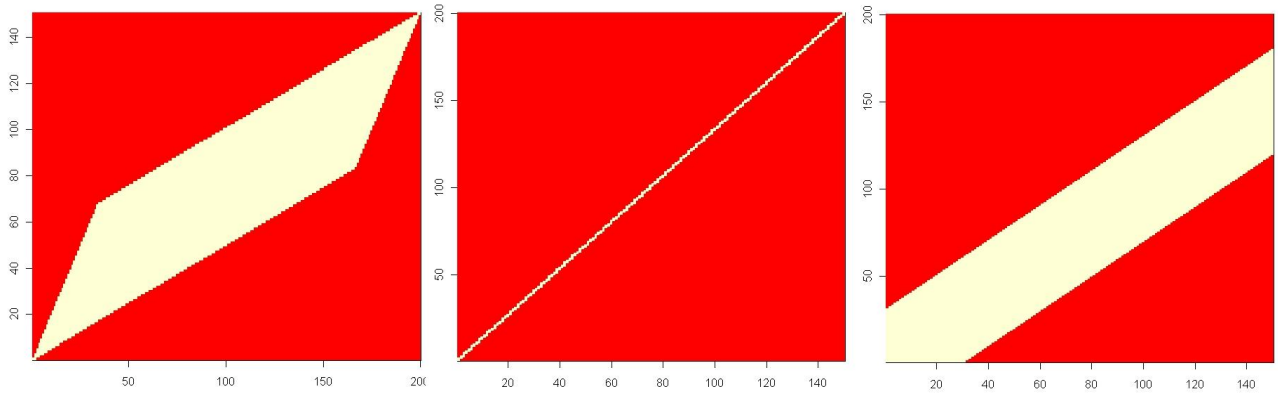


FIGURE 3.4 – *Exemples de fenêtres. De gauche à droite : la fenêtre d'Itakura, la "slanted band", et la fenêtre de Sakoe Chiba.*

2. Modèles de pas :
   Différents pas ont été proposés, qu'ils soient symétriques ou non, autorisés par telle ou telle fenêtre, normalisables ou non. Certains d'entre eux sont présentés sur la figure 3.5, nous ne les détaillerons pas plus ici.

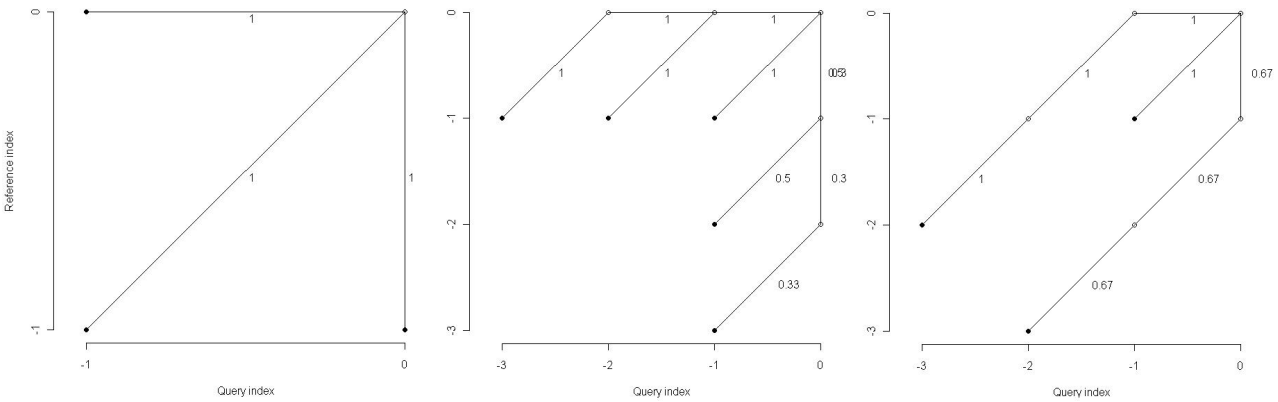


FIGURE 3.5 – *Exemples de modèles de pas*

Ainsi la dtw nous permet d'effectuer la comparaison présentée par la figure 3.6.

Une autre méthode permettant de calculer la distance entre deux vecteurs de tailles différentes revient à les modifier afin qu'ils soient de même taille. L'interpolation permet

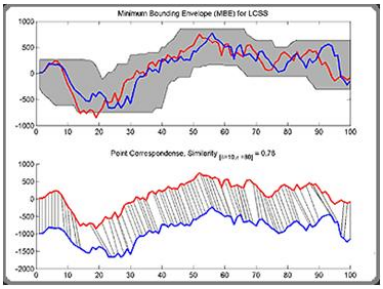


FIGURE 3.6 – *Exemple de Dynamic Time Warping*

d'obtenir des données cohérentes entre deux données déjà connues, et peut donc redimensionner nos trajectoires.

3.2 Interpolation

En analyse numérique (et dans son application algorithmique discrète pour le calcul numérique), l'interpolation est une opération mathématique permettant de construire une courbe à partir de la donnée d'un nombre fini de points, ou une fonction à partir de la donnée d'un nombre fini de valeurs. La solution du problème d'interpolation passe par les points prescrits, et, suivant le type d'interpolation, il lui est demandé de vérifier des propriétés supplémentaires.

L'interpolation doit être distinguée de l'approximation de fonction, qui consiste à chercher la fonction la plus proche possible, selon certains critères, d'une fonction donnée. Dans le cas de l'approximation, il n'est en général plus imposé de passer exactement par les points donnés initialement. Ceci permet de mieux prendre en compte le cas des erreurs de mesure, et c'est ainsi que l'exploitation de données expérimentales pour la recherche de lois empiriques relève plus souvent de la régression linéaire, ou plus généralement de la méthode des moindres carrés.

3.2.1 Interpolation linéaire

Dans le cas d'une interpolation linéaire, on constitue une courbe d'interpolation qui est une succession de segments. Entre deux points p_1 et p_2 de coordonnées respectives (x_1, y_1) et (x_2, y_2), l'interpolation est donnée par la formule suivante :

$y = p.(x - x_1) + y_1$, avec la pente p qui s'exprime comme $p = \frac{y_2 - y_1}{x_2 - x_1}$.

3.2.2 Interpolation cosinus

On utilise ici la fonction cosinus pour modéliser localement la courbe. Deux points seulement sont nécessaires pour évaluer la fonction qui remplace la courbe discrète. La tangente à chaque pic est horizontale, ce qui signifie que chaque pic de la courbe correspond réellement à un point connu de la courbe discrète.

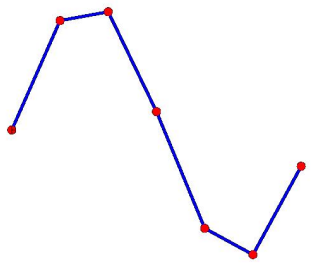


FIGURE 3.7 – *Interpolation linéaire*

3.2.3 Interpolation cubique

Comme son nom l'indique, on utilise ici une équation cubique pour modéliser localement la courbe. Quatre points sont nécessaires pour évaluer la fonction qui remplace la courbe discrète. Tout dépend des conditions de continuité utilisées, la forme de la cubique peut varier et donner une interpolation différente (ex : interpolation cubique de Keys ou interpolation cubique splines). La tangente en chaque point d'indice "i" possède la même pente que le segment reliant les points d'indice "i-1" et "i+1", ce qui signifie que chaque pic de la courbe peut être dépassé par la courbe interpolée.

3.2.4 Interpolation polynomiale

Une interpolation polynomiale consiste à utiliser un polynôme unique (et non des tronçons comme précédemment), de degré aussi grand que nécessaire, pour estimer localement l'équation représentant la courbe afin de déterminer la valeur entre les échantillons.

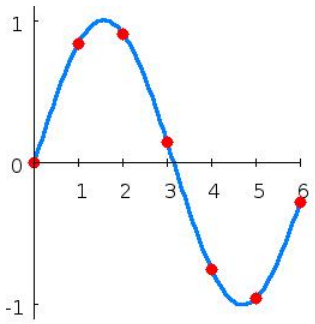


FIGURE 3.8 – *Interpolation polynomiale*

3.2.4.1 Le polynôme de Lagrange

Le polynôme de Lagrange associé au point (x_i, y_i) est :

$$l_i(x) = \frac{x-x_1}{x_i-x_1}...\frac{x-x_{i-1}}{x_i-x_{i-1}}.\frac{x-x_{i+1}}{x_i-x_{i+1}}...\frac{x-x_N}{x_i-x_N}.$$

Les l_i sont de degré $N-1$.

On peut vérifier qu'on a $l_i(x_i) = 1$ et $l_i(x_j) = 0$. Grâce à ces polynômes, nous pouvons interpoler tout ensemble de N points par un polynôme de degré $N > 1$. Et nous avons même plus :

Théorème : Le polynôme $L(x) = y_1.l_1(x) + ... + y_N.l_N(x)$ est l'unique polynôme de degré au plus $N-1$ vérifiant $L(x_i) = y_i$.

Le résultat n'est toutefois pas toujours à la hauteur des espérances : l'interpolation de Lagrange, par exemple, peut fort bien diverger même pour des fonctions très régulières

3.2.5 Le phénomène de Runge

Dans le domaine mathématique de l'analyse numérique, le phénomène de Runge se produit dans certains problèmes d'interpolation de fonctions lorsqu'on augmente le nombre de points d'interpolation. Il montre que cette augmentation ne constitue pas nécessairement une bonne stratégie d'approximation de la fonction.

Par définition, dans un problème à n points d'interpolation, le polynôme d'interpolation doit coïncider avec la fonction en chacun des n points et peut être de degré aussi élevé que $n-1$. Lorsque n augmente, on pourrait s'attendre à ce que fonction et polynôme d'interpolation deviennent de plus en plus proches. Cependant le mathématicien Carle David Tolmé Runge découvrit en étudiant l'erreur d'approximation entre une fonction et ses polynômes interpolateurs qu'ils pouvaient au contraire s'écarter de plus en plus fortement.

Considérons la fonction $f(x) = \frac{1}{1+25x^2}$.

On considère $(n+1)$ points équirépartis dans le segment $[-1, 1]$:

$$x_0 = -1, x_1 = x_0 + h, \cdots, x_{k+1} = x_k + h = x_0 + (k+1)h, \cdots, x_n = 1 \text{ avec } h = \frac{2}{n}.$$

Enfin, on considère le polynôme interpolateur de f aux points x_i, c'est-à-dire l'unique polynôme P de degré inférieur ou égal à n tel que $P(x_i) = f(x_i)$ pour tout i. On note P_n ce polynôme.

Runge a montré que l'erreur d'interpolation entre P_n et f tend vers l'infini lorsque n augmente. Autrement dit, plus on fixe de points où le polynôme a la même valeur que f, moins bien on approche la fonction ! Formellement cela donne :

$$\lim_{n\to\infty} \left(\max_{-1\leq x\leq 1} |f(x) - P_n(x)|\right) = \infty.$$

En fait, lorsqu'on augmente le nombre de points, on constate que le polynôme se met à osciller fortement entre les points x_i avec une amplitude de plus en plus grande, comme l'illustre la figure 3.9.

On peut minimiser l'oscillation des polynômes interpolateurs en utilisant les points de Tchebychev au lieu de points équirépartis pour interpoler. Dans ce cas, on peut montrer que l'erreur d'interpolation (c'est-à-dire $\max_{-1\leq x\leq 1} |f(x) - P_n(x)|$) décroît lorsque n augmente.

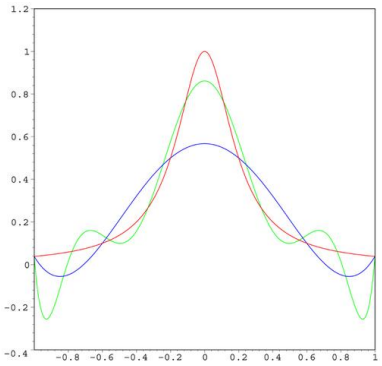


FIGURE 3.9 – *La courbe rouge est la fonction de Runge ; la courbe bleue est le polynôme interpolateur de degré 5 et la courbe verte est le polynôme interpolateur de degré 9. L'approximation est de plus en plus mauvaise.*

Le phénomène de Runge met en lumière le fait que l'interpolation polynomiale n'est pas bien adaptée à l'approximation de fonctions. Pour approcher une fonction avec des polynômes, on peut préférer utiliser des splines par exemple (ce sont des polynômes par morceaux). Dans ce cas, pour améliorer l'approximation, on augmente le nombre de morceaux et non le degré des polynômes.

3.2.6 Splines

Comme on souhaite interpoler par des fonctions plus différentiables que les polynômes ci-dessus présentés , on cherche l'interpolante dans l'espace des fonctions de classe $C^{k>1}$ polynômiales de degré k sur chaque intervalle $[t_i, t_{i+1}]$, appelées fonctions splines (la terminologie a été introduite par Schoenberg en 1946 : en anglais, spline désigne une bande de métal souple utilisée par les dessinateurs pour tracer une jolie courbe entre deux points). C'est ainsi que sont nées les fonctions et les courbes B-splines.

On se limite au courbes de degré 3, pour simplifier. On cherche à faire passer une courbe B-spline de degré k de positions et vitesses aux extrémités prescrites par $N > 1$ points Q_i. Le problème se divise en deux phases.

- Première phase : On se fixe un vecteur de noeuds t et on cherche un polygone de contrôle P tel que la courbe B-spline X_k correspondante passe par les Q_i aux noeuds. L'interpolation se traduit alors par la résolution d'un système linéaire.
- Deuxième phase : on cherche à optimiser le choix du vecteur de noeuds. C'est typiquement non linéaire.

3.2.6.1 Le problème linéaire

On se contente ici d'énoncer le théorème pour des B-splines de degré 3. Une généralisation à tout degré impair se trouve dans :

Théorème : Soient $Q_0, ..., Q_N$ des points de R^n . Soient v_a , v_b deux vecteurs de R^n . Soit t un vecteur de noeuds vissé aux extrémités, de la forme

$t_0 = t_1 = t_2 = t_3 = a < t_4 < ... < t_{N+2} < b = t_{N+3} = ... = t_{N+6}$.

Il existe un unique polygone de contrôle $P = (P_0, ..., P_{N+2})$ tel que la courbe B-spline de degré 3 associée satisfasse :
$\forall j = 0, ..., N, X_3(t_{j+3}) = Q_j, X_3'(a) = v_a$ et $X_3'(b) = v_b$.

Lemme : Soient $f, x : [a, b] \in R$ deux fonctions de classe C^2. On suppose que
– x est polynomiale de degré 3 sur chaque intervalle $[ti, ti + 1], i = 3, ..., N + 2$
– $f(ti) = x(ti)$ pour $i = 3, ..., N + 3$ et $f'(a) = x'(a), f'(b) = x'(b)$.
Alors $\int_a^b (f''(t) < x''(t))^2 dt = \int_a^b f''(t)^2 dt - \int_a^b x''(t)^2 dt$

Remarque : Ce lemme signifie que la spline x qui interpole une fonction f est la projection orthogonale de f sur le sous-espace des splines, pour le produit scalaire $f \Delta g = \int_a^b f''(t) g''(t) dt$

3.2.6.2 Estimation de l'erreur d'interpolation

Théorème : Soit $f : [a, b] \to R$ une fonction de classe C^2. Soit X_3 la fonction B-spline de degré 3 qui l'interpole en $N + 1$ points plus les dérivées aux bornes. Alors
$||f - X_3||_\infty \geq \frac{h^{\frac{3}{2}}}{2} ||f''||_2$ et $||f' - X_3'||_\infty \geq h^{\frac{1}{2}} ||f''||_2$
où $h = max\{|t_{i+1} - t_i|\}$.

3.2.6.3 Choix du vecteur de noeuds

Etant donnés les points Q_i à interpoler, quel est le meilleur choix des t_i ? Pour éviter que la dérivée de la courbe interpolante soit grande, il vaut mieux que des points Q_i et Q_{i+1} éloignés soient interpolés en des valeurs t_i et t_{i+1} éloignées. Autrement dit, il faut corréler les espacements $t_{i+1} < t_i$ avec les distances $||Q_{i+1} - Q_i||$.

Un choix simple consiste à poser $t_i - t_{i+1} = ||Q_{i+1} - Q_i||$.

Ce choix s'appelle la paramétrisation cordale (chordal parametrization). Elle a pour effet de produire une courbe interpolante paramétrée à vitesse à peu près constante.

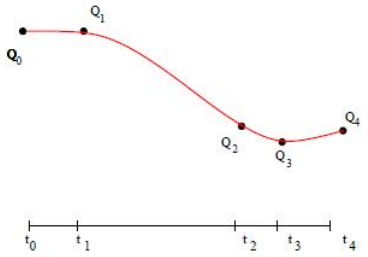


FIGURE 3.10 – *Le choix de noeuds : la paramétrisation cordale*

3.2.7 Méthode des plus proches voisins

Il existe aussi des méthodes dites des plus proches voisins qui consistent à approximer un point par la moyenne d'un nombre choisi de points voisins. Il est possible de pondérer cette moyenne par la distance aux points. Ce qui en pratique, si nous choisissons de considérer

les deux plus proches voisins, nous ramène à une interpolation linéaire entre les deux points les plus proches. La solution sera une fonction affine par morceaux.

Chapitre 4

L'échantillon

4.1 Introduction

Il était initialement prévu que l'observatoire de trajectoires serait installé pendant plusieurs mois sur un même site, mais il s'est avéré que cette expérience devait être reportée. Les mesures n'ont donc pas été disponibles dans le cadre de ce stage et c'est sur un échantillon de mesures de l'OdT prises les 23 et 25 juin 2009 que nous avons travaillé. Notre travail fut dans un premier temps de nettoyer ces données, car l'OdT n'était pas encore sous sa forme finale.

4.2 Filtres

Même si l'Observatoire de trajectoire est à ce jour l'un des instruments les plus précis pour décrire une trajectoire, ses mesures peuvent présenter des incohérences, nous tâcherons ici de sélectionner les trajectoires viables. Pour commencer, nous disposons d'un échantillon de 3007 trajectoires allant toutes dans la même direction, sauf 5 d'entre elles. Ces dernières n'empruntent pas la même voie, et tournent vers la droite, tandis que les autres tournent vers la gauche, nous les supprimons donc.

La figure 4.1, met en évidence que toutes les trajectoires ne commencent ni ne finissent au même point. Il nous faut choisir la zone que nous allons étudier. Pour ce faire, observons le marquage central, et plus précisément son *rayon de courbure*.

Définition du rayon de courbure

Prenez une courbe, et prenez un point A sur cette courbe. Tracez la normale à cette courbe, et prenez un point O sur la normale. Alors, le cercle de centre O passant par A est tangent à la courbe. Mais tous les cercles tangents à la courbe ne sont pas tangents de la même façon... En effet, si O est proche de A, le cercle va se situer plutôt "à l'intérieur de la courbe". Si O est loin de A, le cercle sera plutôt "à l'extérieur de la courbe". Le rayon limite entre être "à l'intérieur de la courbe" et être "à l'extérieur de la courbe" s'appelle le rayon de courbure de la courbe au point A. Le cercle correspondant se nomme le cercle osculateur cf 4.2. Si la courbe est la courbe représentative d'une fonction f, donnée donc par (x,f(x)), on a :

$$R = \frac{(1+f'^2)^{3/2}}{f''}.$$

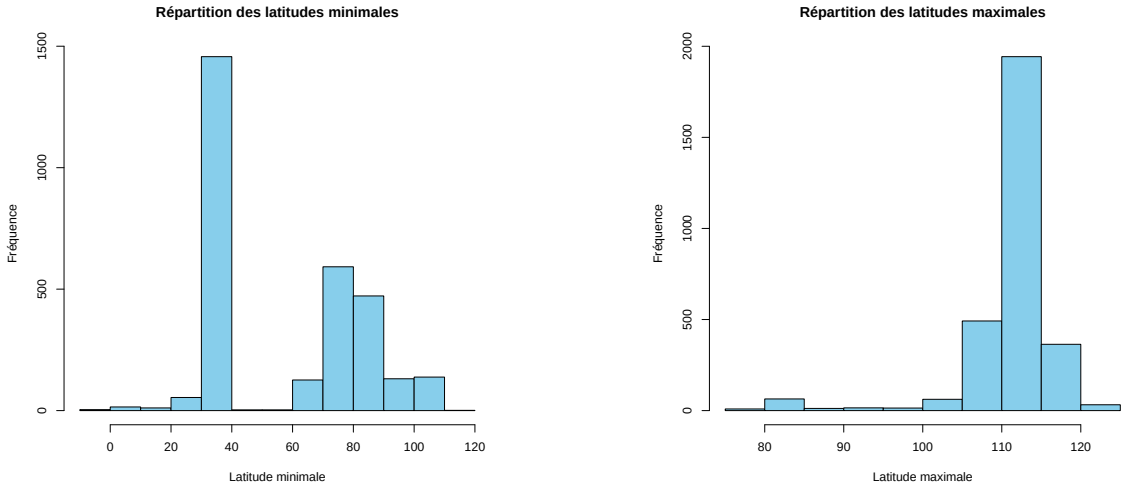


FIGURE 4.1 – *Répartition des latitudes minimales (à gauche) et maximales (à droite)*

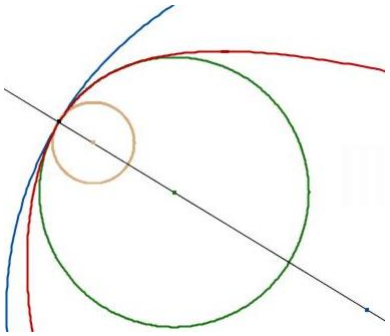


FIGURE 4.2 – *Le rayon de courbure*

La fonction Spline de R permet d'obtenir l'interpolée par spline d'une fonction, ainsi que ses dérivées. En l'appliquant à la fonction liant la *latitude* du marquage central à sa *longitude*, nous pouvons obtenir son *rayon de courbure* affiché sur la figure 4.3. Le *rayon de courbure* commence à diminuer de façon significative à partir de 20 en abscisse curviligne, mais seulement 41% des trajectoires sont décrites à cet endroit ; de même seulement 13% des trajectoires sont encore décrites quand le marquage central repars en ligne droite. Nous allons donc restreindre la zone d'étude à $26 - 96$ en abscisse curviligne du marquage central. Ceci correspond, d'après le tableau 4.1 à $45 - 109$ en *Latitude*.

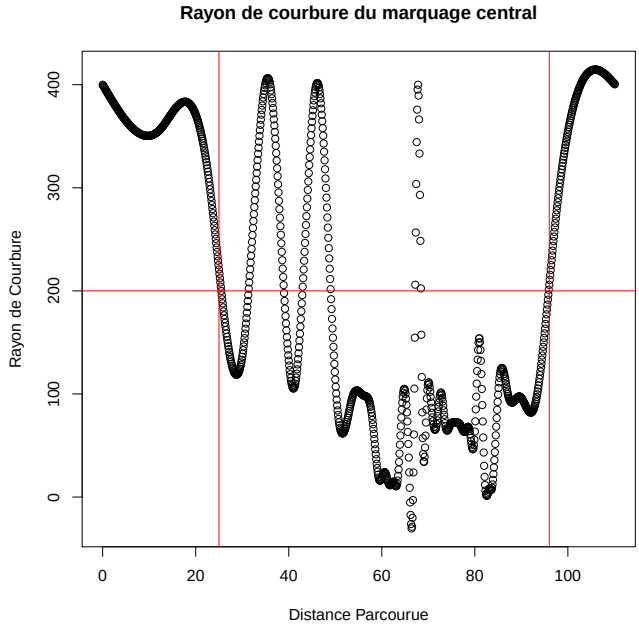


FIGURE 4.3 – *Rayon de courbure du marquage central*

Latitude	Longitude	Abscisse curviligne	Rayon de courbure
45.92	176.58	26	280.13
109.07	203.37	96	205.85

TABLE 4.1 – *Zone d'étude*

4.2.1 Premier filtre : La zone d'étude

Supprimons dans un premier temps, les trajectoires dont la *latitude* minimale est inférieure à la *latitude* minimale exigée (45). Ces trajectoires, par définition, ne recouvrent pas la zone d'étude. Nous en trouvons 1463, soit 49% de la base initiale.

Supprimons à présent les trajectoires dont la *latitude* maximale est supérieure à la *latitude* maximale exigée (109). Ces trajectoires, par définition, ne recouvrent pas la zone d'étude. Nous en trouvons 476, soit 16% de la base initiale.

Certaines trajectoires, parmi celles que nous avons conservées, sont décrites par trop peu d'observations pour qu'une interpolation nous permette ensuite de les étudier correctement. Arbitrairement, nous décidons qu'un écart de 10m, soit environ 14% de la zone d'étude, entre deux mesures est trop important. Ces trajectoires, que nous supprimons de la base initiale, sont au nombre de 81 soit 3% des trajectoires initiales.

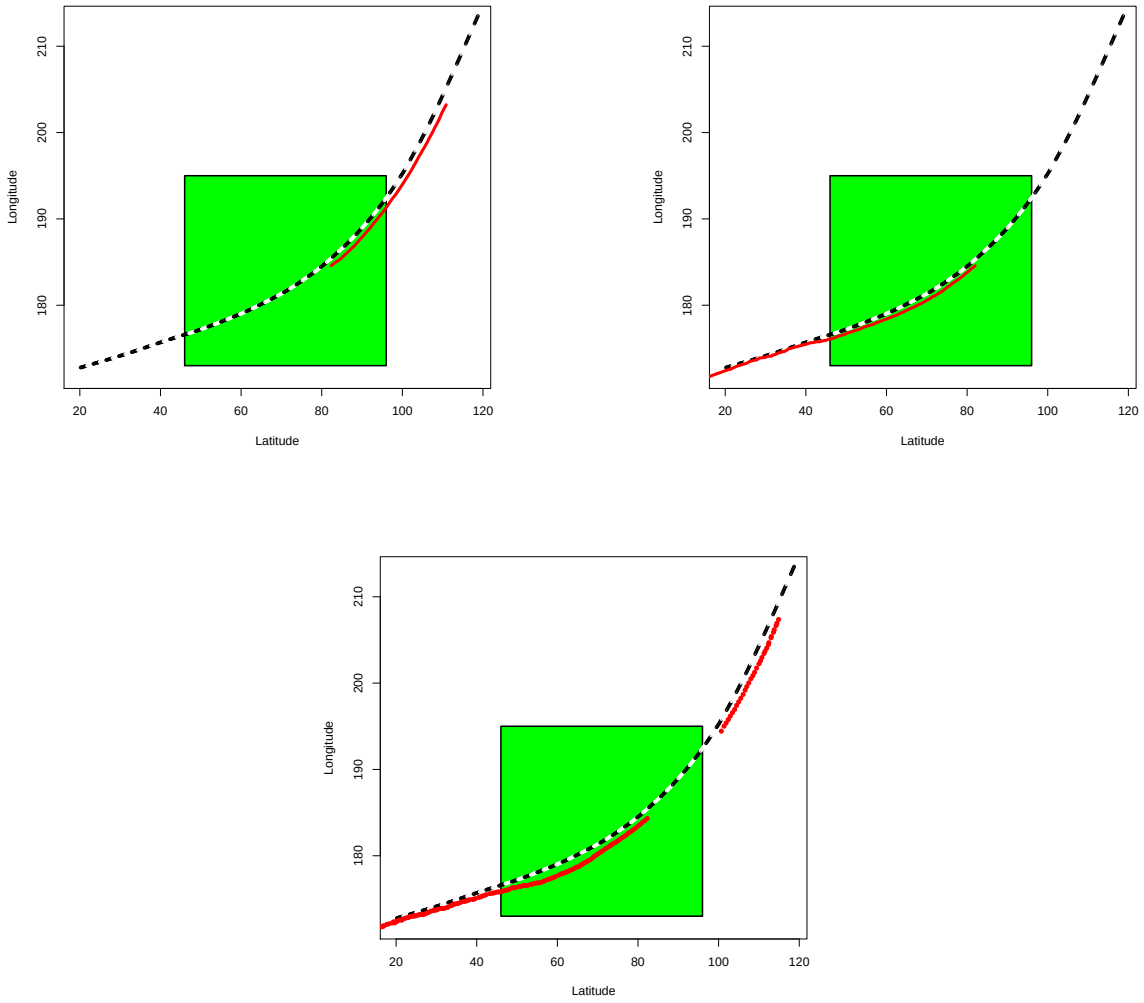


FIGURE 4.4 – *Exemples de trajectoires supprimées par le premier filtre(les critères sont, de gauche à droite et de haut en bas : latitude minimale, latitude maximale, écart entre les observations)*

Critère	Avant	Après	Supprimées	Ratio
Latitude minimale	3007	1544	1463	49%
Latitude maximale	3007	2531	476	16%
Ecart entre observations	3007	2926	81	3%
Total	3007	1355	1652	55%

TABLE 4.2 – *Premier filtre : La zone d'étude*

4.2.2 Deuxième filtre : défaut de synchronisation

Nous considérerons ici les trajectoires ayant passé le premier filtre. Ces trajectoires sont classées par rapport au *temps*, ce qui n'implique pas que les trajectoires soient de *latitude* croissante. Comme tous les véhicules vont dans le même sens, ceci correspond à des erreurs de mesure (nous mettons ce phénomène en évidence sur la figure 4.5). Celles-ci semblent provenir d'un défaut de synchronisation entre les capteurs. Supprimons donc les observations aberrantes, c'est à dire celles qui font "reculer" le véhicule (la *latitude* à l'instant i+1 est inférieure à la *latitude* à l'instant i) ou celles qui font "sauter" le véhicule (il y a un écart supérieur à 10m entre la *latitude* à l'instant i et la *latitude* à l'instant i+1).

Pour trouver ces données et uniquement ces données, nous utiliserons l'algorithme suivant :

```
latitudeprécédente <- latitude (1)
pour (i parcourant une trajectoire triée par rapport au temps)
si (latitude précédente > latitude(i)) ou si (latitude(i) - latitudeprécédente > 10)
    supprimer observation i
sinon (latitudeprécédente <- latitude i)
```

Ce qui permet de supprimer les observations aberrantes, et uniquement celles-ci.

Lorsqu'on applique les filtres spatiaux (section 4.2.1) aux données allégées des observations aberrantes décrites plus haut, il ne nous reste aucune trajectoire. Ces données aberrantes seraient dues à une mauvaise synchronisation entre les capteurs, comme le montre la figure 4.6.

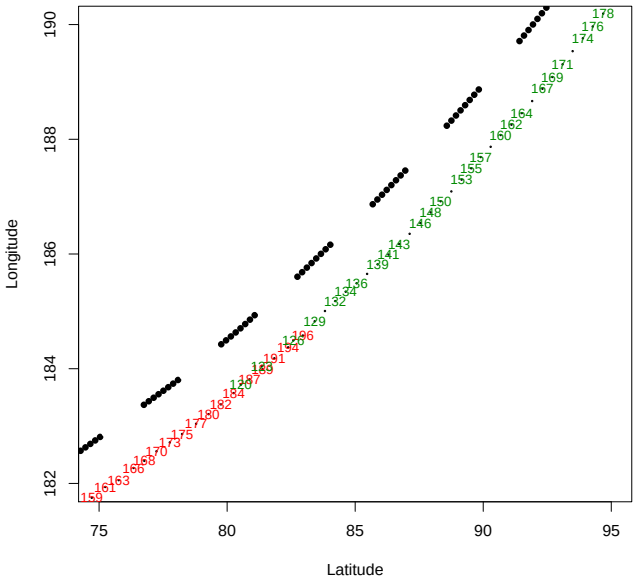


FIGURE 4.5 – *Exemples de sauts et de reculs : En vert les observations suivant un "saut", en rouge les observations suivant un "recul". les numéros correspondent à l'ordre des données triées par rapport au temps. En observant par exemple l'indice 160, on retrouve le 161 bien en retrait, puis le 162 près du 160, le 163 près du 161, et le 164 après le 162, le 165 respectant l'ordre, il est représenté par un simple point.*

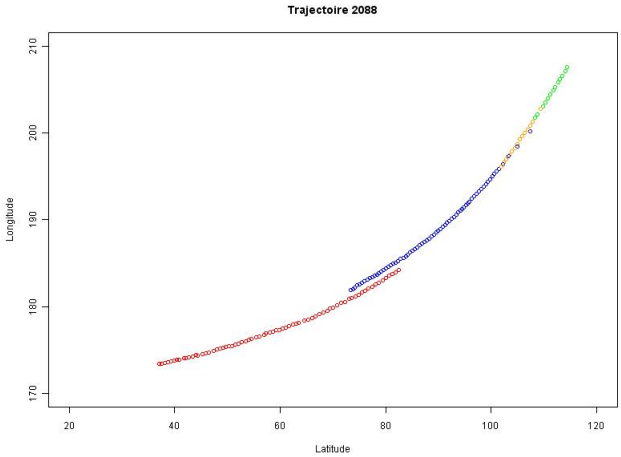


FIGURE 4.6 – *Exemple de trajectoire : en rouge, les observations de la première caméra, en bleu : les observations du télémètre, en jaune : la fusion des données du télémètre et de la seconde caméra, en vert, les données de la seconde caméra.*

4.2.3 Troisième filtre : Sélection des capteurs

Nous avons vu qu'une sélection brute des données ne laissait aucune trajectoire cohérente, il nous faut donc trouver un autre critère de sélection. Nous avons observé sur la figure 4.6 que certaines aberrations était dues à une mauvaise synchronisation des capteurs. Ne retenir les mesures d'un seul capteur nous permettrait donc d'éviter ces aberrations. Nous lisons dans la thèse de Yann Goyat (Goyat [2008]), concepteur de l'OdT : "Les données les plus riches pour l'analyse de trajectoire étant situées au centre des virages, il a été décidé de donner un poids particulier à cette zone de mesure." Le télémètre étant "très stable", il couvre cette zone. Pour ces raisons, nous retiendrons les mesures du télémètre, correspondant au *champ* 2.

Toutes les trajectoires ne sont pas observées par le télémètre (seulement 2763 trajectoires ont au moins une observation correspondant au *champ* 2). Il nous faut à présent définir une nouvelle zone d'étude, plus petite que la précédente.

Nous nous proposons tout d'abord de décrire les *latitudes* recouvertes par les trajectoires : la figure 4.7 nous enseigne à ce sujet que la *latitude* minimale sera entre 70 et 90 , et que la *latitude* maximale sera autour de 110. La figure 4.8 nous montre l'évolution du nombre de données en fonction de l'étendue des *latitudes* exigées. Afin de conserver un nombre conséquent de trajectoire recouvrant une distance suffisante, nous choisissons de prendre les trajectoires allant de 78 à 108 en *latitude*, zone représentée sur la figure 4.9. Les trajectoires appartenant à cette zone vérifient les critères du premier filtre appliqués à ces nouvelles limites.

FIGURE 4.7 – *histogrammes des latitudes initiales (à gauche) et finales(à droite)*

Critère	Avant	Après	Supprimées	Ratio
Zone d'étude	2763	1168	1595	58%
Ecart entre observations	1168	1131	37	3%
Total	2763	1131	1632	59%

TABLE 4.3 – *Premier filtre appliqué aux données du télémètre : La nouvelle zone d'étude*

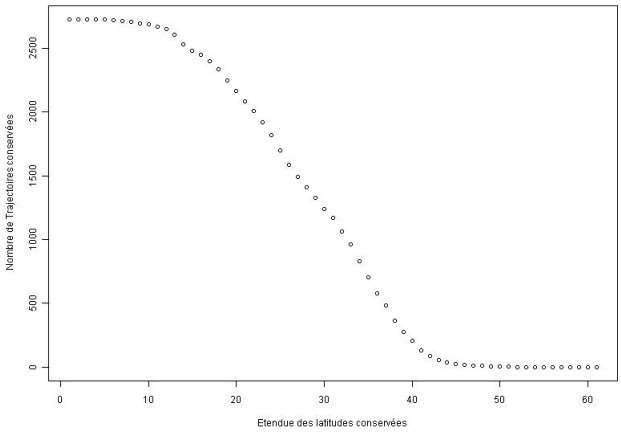


FIGURE 4.8 – *Nombre maximum de trajectoires conservées en fonction de l'étendue des latitudes. Nous l'obtenons en prenant le maximum des intervalles de même longueur.*

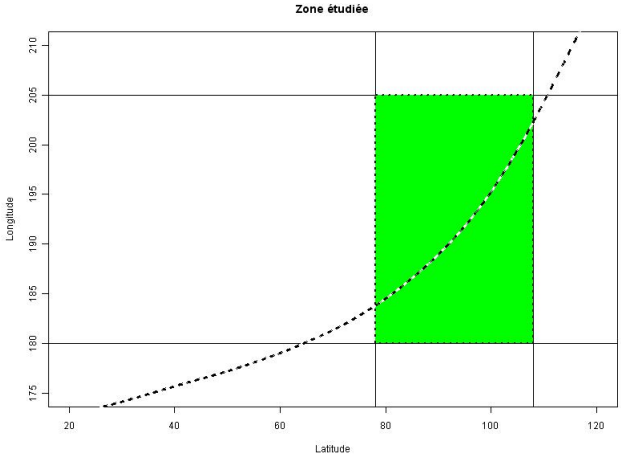


FIGURE 4.9 – *Nouvelle zone d'étude*

4.2.4 Quatrième filtre : filtre des variables.

Reprenons l'ensemble des variables qui nous sont proposées initialement, et tâchons de sélectionner les plus pertinentes.

Certaines variables nous semblent inutiles :

- *Date Du Fichier, Année, Mois, Jour, Heure, Minute, Seconde* : toutes les trajectoires étant prises le même jour, ces données nous sont inutiles.
- *Longueur,Largeur et Hauteur* du véhicule : Nous nous pencherons surtout sur les trajectoires. Certains filtres précédents nous laissant penser que plusieurs trajectoires aient pu être confondues, nous nous étonnons de trouver ces variables constantes par

trajectoire.
- *Type* : Les trois types dont nous disposons sont clair, sombre, et indéfini. Notre étude ne cherchant pas à étudier le type de conduite en fonction de la couleur du véhicule, nous ne prendrons pas cette variable en compte.
- *Champs* : Nous avons choisi le champ 2, correspondant aux données mesurées par le télémètre.
- *Sens* : Comme nous avons supprimé au départ les 5 véhicules allant en sens inverses, cette variable est identique pour toutes les trajectoires.

FIGURE 4.10 – *Exemple de trajectoire (à gauche) dont l'écart par rapport au marquage central est clairement faux : le véhicule traverse le marquage central mais l'écart reste positif.*

D'autres variables ne sont pas viables :
- *Ecart* : la figure 4.10 montre bien que des trajectoires qui traversent le marquage central peuvent ne pas avoir d'écart négatif. Cette variable n'est donc pas suffisamment précise pour être utilisée dans le cadre d'une analyse statistique, nous la calculerons donc à notre tour, comme ceci :
  pour chaque point de la trajectoire, il nous faut calculer la distance qui le sépare du marquage central. Plusieurs solutions s'offrent à nous : nous pouvons, par exemple, prendre la distance entre le point et son projeté sur le marquage central ou prendre la distance minimale séparant ce point des points composant le marquage central. C'est cette dernière solution que nous retiendrons.
- *Position* : dépendant de l'écart, elle n'est pas non plus valable.
- *Angle Au Volant* : certaines trajectoires ne se retrouvant pas du tout dans l'angle au volant leur correspondant, cette variable n'est pas valable.

Certaines, enfin, sont utilisables :
- *Numéro.*
- *Temps,Intervalle de Temps.*
- *Latitude,Longitude.*
- *Vitesse.*
- *Cap.*

Auxquelles nous ajoutons :
- *diffecart* : la différence entre les écarts successifs, qui prend le rôle de l'*angle au volant* en traduisant la direction du véhicule.
- *Accélération* : la dérivée de la *vitesse* par rapport au temps.
- *Le rayon de courbure* qui remplit le même rôle que *diffecart*, à savoir traduire la direction du véhicule.
- *L'accélération latérale*, détaillée ci-après.

Définition de l'accélération latérale : L'accélération latérale d'un véhicule définit, comme son nom l'indique, l'accélération du véhicule vers ses côtés. Elle représente la poussée qui s'oppose au virage d'un véhicule : plus le véhicule va vite, plus l'accélération latérale est forte. Son équation est la suivante :

$$\frac{V^2}{R}$$

Pour calculer le *rayon de courbure*, nous effectuons d'abord un lissage par splines. Il permet d'obtenir une courbe sans valeurs exotiques, le *rayon de courbure* réagissant très vite aux variations. Puis la fonction 'splinefun' de R nous renvoie les dérivées de la trajectoires. Cette méthode nous permet de repérer une trajectoire qui a deux *longitudes* différentes pour une même *latitude*. Nous supposons alors que deux trajectoires y ont été confondues. Nous supprimons donc cette trajectoire, ce qui réduit notre base finale à 1130 trajectoires.

Nous recalculons également la variable *position* à partir des nouveaux *écarts* : ne connaissant pas la largeur de la route, nous ne pouvons que nous baser sur les *écarts* dont nous disposons pour définir les sept *positions* différentes. Ainsi, en prenant $max = max(ecarts)$ et $min = min(ecarts)$, et $pas = (max - min)/7$, nous obtenons les 7 intervalles $[min; min + pas]$ ... $[max - pas; max]$ correspondant aux *positions* 6, 5 ... ,0.

4.3 Jeu de données final

Nous avions initialement 23 variables qui décrivaient 3007 trajectoires d'une latitude minimale égale à -8 à une latitude maximale égale à 127. Nos filtres nous ont montré que nous ne pouvions nous fier qu'à 14 variables (tableau 4.4) décrivant 1130 trajectoires recouvrant des latitudes allant de 78 à 108.

Variable	Desciption
Numéro	Identifiant du trajet.
Distance parcourue	Distance parcourue par le véhicule depuis la première mesure retenue (en m).
Temps écoulé	Temps écoulé depuis la première mesure retenue (en 10^{-6} secondes).
Latitude	Latitude du véhicule, un offset lui est appliqué afin de rendre les valeurs plus maniables (en m).
Longitude	Longitude du véhicule, un offset lui est appliqué afin de rendre les valeurs plus maniables (en m).
Vitesse	Vitesse du véhicule, calculé par le SAVe (en km/h).
Accélération	Dérivée de la vitesse par rapport au temps écoulé (en m/s^2).
Rayon de courbure	Rayon de courbure de la trajectoire en chacun de ses points (en m).
Accélération latérale	Accélération latérale du véhicule obtenue à partir de la vitesse et du rayon de courbure (en m/s^2).
Cap	Cap pris par le véhicule (direction).
Angle Au Volant	Calculé à partir des Caps successifs.
Ecart	Ecart au marquage central (en m).
DiffEcart	Différence entre les écarts successifs (en m).
Position	Position du véhicule, de 0 pour les véhicules, près du marquage central, à 6 pour les véhicules loin du marquage central.

TABLE 4.4 – *Liste finale des variables*

Chapitre 5

Analyse des données

Notre étude porte sur le jeu de données final présenté à la section 4.3. Elle progressera à travers trois parties : l'analyse descriptive des résultats, l'analyse par indicateurs et la proposition d'un indice de risque fondé sur l'écart. Par l'analyse descriptive, nous observerons l'allure générale des variables retenues (tableau 4.4). L'analyse par indicateurs impliquera dans un premier temps la création de nombreux indicateurs, puis la sélection des indicateurs les plus pertinents pour la classification des trajectoires. Enfin, nous construirons un indice de risque, fonction de la variable *écart*, dont nous tâcherons d'évaluer la qualité à partir des indicateurs sélectionnés dans la section précédente.

5.1 Analyse descriptive

Parmi les variables listées dans la table 4.4, on distingue deux groupes : les variables décrivant la position du véhicule (*latitude*, *longitude*, *écart*...), et celles décrivant sa dynamique (*vitesse*,*accélération*...). L'*écart* fournit plus d'information que la *latitude* et la *longitude*, il nous explique en effet le comportement du véhicule par rapport au marquage central, si l'on observe également la *vitesse*, nous disposerons des principaux éléments décrivant une trajectoire.
Nous observons sur la figure 5.1 que les *vitesses* instantanées suivent une distribution normale de moyenne 62.4 et d'écart type 7.7. Aucun véhicule ne roule à moins de 30 km/h, ni aucun véhicule n'excède 95 km/h. Nous disposons donc d'un échantillon de *vitesses* varié, sans *vitesse* aberrante. Les filtres ont ainsi permis de garder uniquement des *vitesses* cohérentes.
Cette même figure nous apprend que l'*accélération* est symétrique autour d'une moyenne nulle, son écart type est très faible (inférieur à 10^{-6}). La tendance générale est donc de conserver une même *vitesse* avec une faible dispersion.
D'après ce même graphique (figure 5.1) les *accélérations latérales* se situent globalement, d'après leur boîte à moustaches, entre $-0.5m^2/s$ et $2m^2/s$. Les véhicules suivent donc bien le virage, mais il faut souligner qu'il ne vont pas constamment vers l'intérieur alors que nous sommes ici en plein coeur du virage. Nous disposons encore de valeurs singulières, qu'il sera intéressant d'étudier : une *accélération latérale* supérieure à 3 n'est pas commune, tout comme une *accélération latérale* supérieure à 1 en valeur absolue, dans le sens opposé au virage.
La variable *DiffEcart*, représentée en figure 5.1 par sa boîte à moustaches, met en évidence

que les écarts au centre de la voie restent globalement constants. Certaines valeurs, qu'il sera intéressant d'analyser plus en détail, sortent une nouvelle fois du lot : bien que notre échantillon ne regorge pas de données, il semble offrir de vraies perspectives à notre analyse.

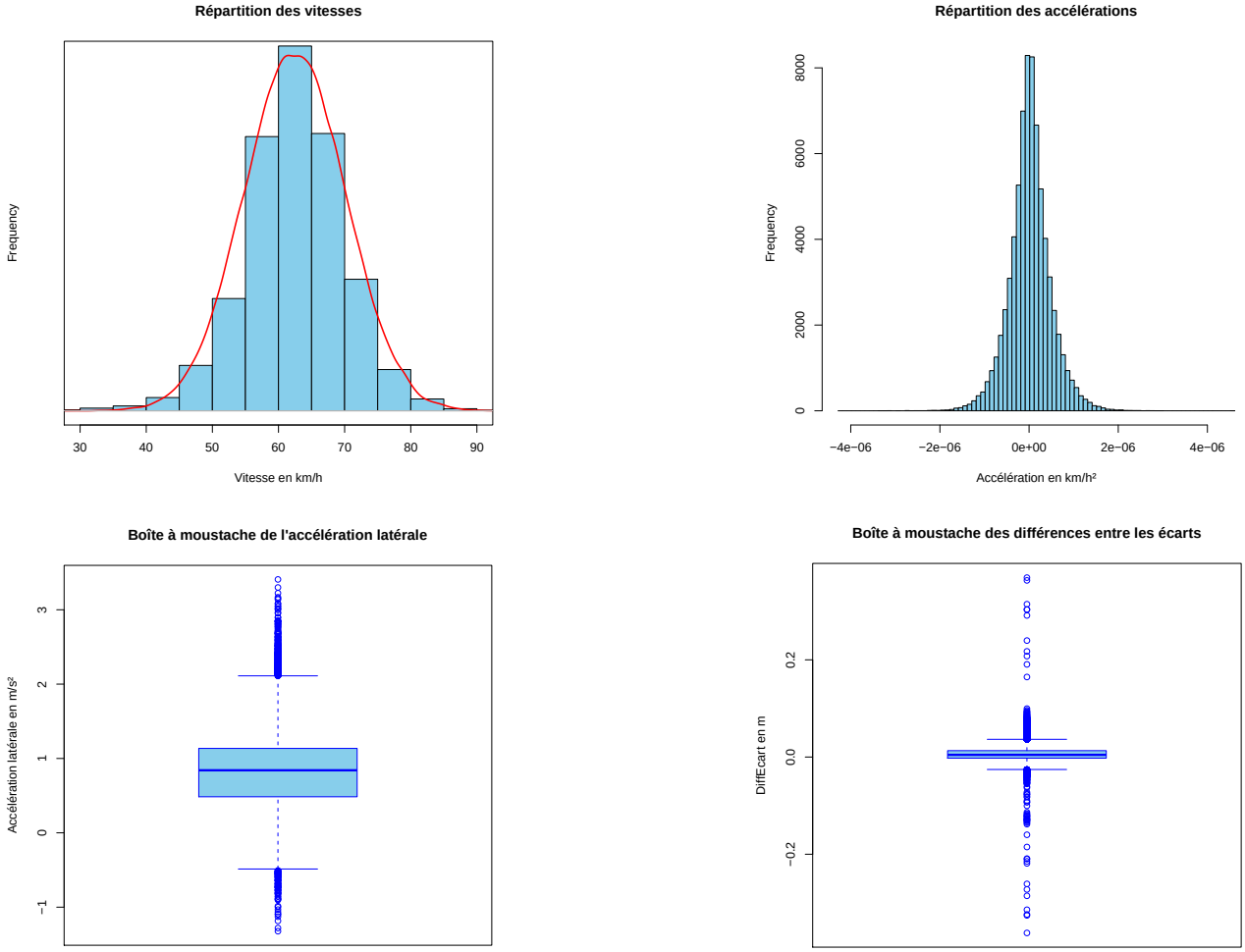


FIGURE 5.1 – *Analyse descriptive : de gauche à droite et de haut en bas, répartition des vitesses instantanées mesurées, répartition des accélération, boîte à moustaches de l'accélération latérale, et boîte à moustache des différences entre écarts.*

La figure 5.2 présente la *vitesse* moyenne des véhicules par *position* par rapport au marquage central (rappelons que 0 signifie à l'extérieur, et que 6 signifie au centre). La plupart des véhicules circulent au centre de la voie (positions 3,4,5), très peu d'entre eux s'aventurant vers l'extérieur. Les *vitesses* correspondent aux résultats des précédentes analyses (Goyat et al. [2008a]) : les véhicules rapides sont plutôt à l'intérieur du virage tandis que les véhicules lents sont à l'extérieur.

Les variables *Cap* et *Rayon de courbure* semblent difficiles à utiliser : la variable *rayon de courbure* tend vers l'infini lorsque la trajectoire est rectiligne, de plus, elle doit pouvoir

être remplacée par la variable *DiffEcart* ; le véhicule tournant progressivement jusqu'à la fin du virage, la variable *Cap* est strictement croissante et à peu près linéaire par rapport au temps. Le *cap* initial et le *cap* final pourraient nous renseigner sur la direction prise par le véhicule avant et après la zone d'étude, fonction que ne peut remplir *DiffEcart*, car le *cap* se base sur l'orientation illustré par la figure 2.4.

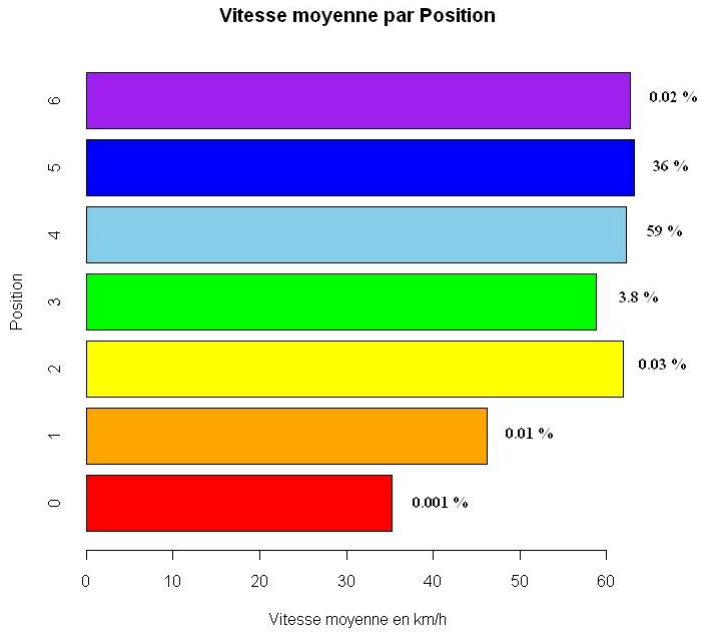


FIGURE 5.2 – *Vitesse moyenne par position, à droite est indiquée le pourcentage de données par position.*

5.2 Analyse par indicateurs

Une trajectoire étant décrite par des mesures successives de multiples variables, il est compliqué de comparer plusieurs trajectoires. L'idée de cette section est de réduire le nombre de variable expliquant une trajectoire à l'essentiel : nous chercherons d'abord à expliquer au mieux les variables par un maximum d'indicateurs, puis à sélectionner les indicateurs les plus pertinents.

5.2.1 Création des indicateurs

5.2.1.1 Calcul des indicateurs

Les indicateurs que nous proposons sont fondés sur les variables retenues au chapitre précédent : *Numéro, Temps, Intervalle de Temps, Latitude, Longitude, Vitesse, Cap, diffecart, accélération, rayon de courbure, accélération latérale*. Pour *Vitesse, Cap, diffecart, accélération, rayon de courbure, accélération latérale* nous calculons la moyenne, le maxi-

mum, le minimum et l'écart type (mean, max, min, std) ainsi que leurs valeurs initiales et finales, nous ajoutons le *PKE* et le *RPA*, décrits ci-dessous.

– Positive Kinetic Energy (*PKE*) : Il est défini de la manière suivante :

$$\frac{\sum\left(v_f^2 - v_s^2\right)}{x},$$

avec, lorsque $dv/dt > 0$, v_f = vitesse finale, v_s = vitesse initiale, x = distance.
Plus l'accélération du véhicule est constante, plus *PKE* est proche de 0. Cet indicateur indique en fait la capacité à garder l'énergie cinétique du véhicule, c'est-à-dire l'élan que possède le véhicule lorsqu'il est en mouvement. Ainsi une conduite nerveuse avec de nombreuses accélérations sera associée à un *PKE* élevé, et inversement une conduite plus fluide sera associée à un *PKE* proche de 0.

– Relative Positive Acceleration (*RPA*) : Il est défini de la manière suivante :

$$\frac{1}{x}\int va^{+}dt,$$

avec x = durée totale du trajet, v = vitesse du véhicule, a = accélération du véhicule.
Cet indicateur augmente lorsque la conduite est caractérisée par de nombreuses accélérations.

5.2.1.2 Liste des indicateurs retenus

Un diagramme de dispersion nous apprend que les variables *rayon de courbure* et *diffecart* ne sont pas tout à fait liées, il nous faut donc en choisir une, qui sera *diffecart*, laquelle nous semble plus fiable.

Si nous avions disposé des trajectoires complètes, nous aurions calculé ces indicateurs sur les trois parties qui composent le virage : l'entrée, le centre et la sortie. Afin de délimiter ces trois parties, plusieurs solutions s'offraient à nous, comme, par exemple de calculer ces indicateurs sur de parties plus petites et plus nombreuses, et d'en découvrir les corrélations. en les classant en trois groupes.

La liste des indicateurs retenues est présentée dans le tableau 5.1.

5.2.2 Classification hiérarchique des variables

5.2.2.1 La procédure VARCLUS du logiciel SAS

Les méthodes hiérarchiques sont des méthodes de classification non supervisées, qui produisent des suites de partitions emboîtées, représentées par un dendrogramme. Cela permet de visualiser toutes les partitions, et d'en choisir la meilleure. On trouve parmi elles les méthodes basées sur un algorithme agglomératif (Classification Ascendante Hiérarchique), et les méthodes basées sur un algorithme divisif (Classification Descendante Hiérarchique).

La procédure "VARCLUS" de SAS implémente une méthode de classification hiérarchique descendante, nous avons trouvé sa description dans Nakache [2004].

La procédure VARCLUS du logiciel SAS conduit à une partition d'un ensemble de variables numériques (quantitatives) en classes disjointes, à partir de la matrice des corrélations (dans ce cas les variables ont le même poids) ou de la matrice des variances covariances (si les variables doivent avoir d'autant plus d'importance dans l'analyse que leurs variances sont grandes).

La partition obtenue est telle que les variables d'une même classe sont aussi corrélées entre elles que possible et deux variables quelconques de deux classes différentes sont le moins corrélées possible.

Variable	Signification
PKE	Capacité à garder l'énergie cinétique (en m/s^2).
RPA	Capacité à garder l'énergie cinétique (en m^2/s^3).
MeanVitesse	Vitesse moyenne (en km/h).
MeanAcceleration	Accélération moyenne (en m/s^2).
MeanCap	Cap moyen.
MeanEcart	Ecart moyen (en m).
MeanDiffEcart	DiffEcart moyen. Il définit la tendance à se rapprocher du marquage central ou à s'en éloigner (en m).
StdVitesse	Ecart type de la vitesse (en km/h).
StdAcceleration	Ecart type de l'accélération (en m/s^2).
StdCap	Ecart type du cap.
StdEcart	Ecart type de la variable écart (en m).
StdDiffEcart	Ecart type de la variable Diffécart (en m).
MaxVitesse	Vitesse maximale (en km/h).
MaxAcceleration	Accélération maximale (en m/s^2).
MaxCap	Cap maximal.
MaxEcart	Valeur maximale de la variable écart (en m).
MaxDiffEcart	Maximum de la variable DiffEcart (en m).
MinVitesse	Vitesse minimale (en km/h).
MinAcceleration	Accélération minimale (en m/s^2).
MinCap	Cap minimal.
MinEcart	Valeur minimale de la variable écart (en m).
MinDiffEcart	Valeur minimale de la variable DiffEcart (en m).
Vitesse Initiale	Première mesure de la vitesse (en km/h).
Ecart Initial	Première mesure de l'écart (en m).
Cap Initial	Première mesure du cap.
Vitesse Finale	Dernière mesure de la vitesse (en km/h).
Ecart Final	Dernière mesure de l'écart (en m).
Cap.Final	Dernière mesure du cap.

TABLE 5.1 – *Indicateurs proposés.*

VARCLUS peut donc être utilisée comme méthode de réduction d'un ensemble de variables de taille importante. Les classes étant construites, on ne garde qu'une variable par classe, celle qui représente au mieux la classe, ramenant ainsi à k (nombre de classes) le nombre p de variables initiales. Souvent en pratique, on contruit k nouvelles variables, cha-

cune d'entre elles représentant une combinaison linéaire des variables qui forment une classe. L'algorithme de VARCLUS, basé sur l'analyse en composantes principales obliques et qui conduit à un système d'axes principaux non orthogonaux est présenté ici de façon succincte :

1. Au départ l'ensemble des variables qui constitue une seule classe notée C est soumis à une ACP et les composantes principales obtenues par rotation orthoblique suivant le critère quartimax et correspondant aux deux plus grandes valeurs propres λ_1 et λ_2 sont retenues si la deuxième plus grande valeur propre λ_2 est supérieur à 1, satisfaisant ainsi le critère le plus simple et le plus populaire connu sous le nom de critère de Kaiser. Chaque variable est alors affectée à la composante principale avec laquelle elle présente le plus fort coefficient de corrélation R^2. Les variables sont itérativement ré-affectées aux classes en maximisant la variance expliquée par les composantes principales. Cette ré-affectation est nécessaire pour maintenir une structure hiérarchique. On obtient ainsi deux classes de variables : C_1 qui contient les variables plus corrélées (en terme R^2) avec la première composante principale qu'avec la deuxième, et C_2 qui contient les variables plus corrélées (en terme R^2) avec la deuxième composante principale qu'avec la première. Chacune de ces deux classes C_1 et C_2 est, à son tour, divisée en deux classes si la deuxième plus grande valeur propre de l'analyse en composantes principales obliques correspondante est plus grande que 1, et ainsi de suite.
2. On effectue une ACP à partir des variables du groupe C_1. Si la deuxième valeur propre de cette ACP est inférieure à 1, le groupe C_1 n'est pas divisé. Dans le cas contraire (critère de Kaiser satisfait par la deuxième plus grande valeur propre), on divise C_1 en deux groupes C_{11} et C_{12} après avoir effectué une rotation oblique des deux premiers axes de cette ACP suivant le critère quartimax comme en (1). On a donc à ce pas de la procédure, trois groupes de variables C_{11}, C_{12}, et C_2. Les variables de ces derniers groupes sont alors ré-affectées à ces groupes de manière à rendre maximum la variance expliquée par la première composante principale de ces groupes.
3. On répète (2) sur C_2 et ainsi de suite.

Les classes de variables sont d'autant plus distinctes (indépendantes) que les R^2 (own cluster) sont grands et les R^2 (next closest) sont petits. Une petite valeur du rapport $(1-R^2$ ratio) indique une bonne typologie des variables. Ce rapport peut prendre pour une variable, une valeur plus grande que 1 si sa corrélation avec sa propre classe est faible et plus petite que sa corrélation avec la classe voisine.

Nous choisissons d'utiliser pour cette classification la matrice des corrélations, car rien ne nous porte à croire que les indicateurs puissent être pondérés.

5.2.2.2 Les classes

On remarque sur la figure 5.3 que ces classes sont homogènes, si l'on occulte les variables "Cap Initial" et "Cap Final". Ces variables ne semblent pas corrélées avec les autres de façon significative, et n'apportent pas assez d'information pour former une classe à elles seules. Elles seront donc supprimées par la sélection qui va suivre : chaque classe sera représentée par la variable l'expliquant le mieux, et la moins corrélée aux autres classes. Ces deux critères ne pointant pas toujours sur le même résultat, certains choix seront motivés par notre connaissance de l'expérience. Nous observons que trois de ces classes sont liées à la position géométrique, tandis que les trois autres sont liées à la vitesse.

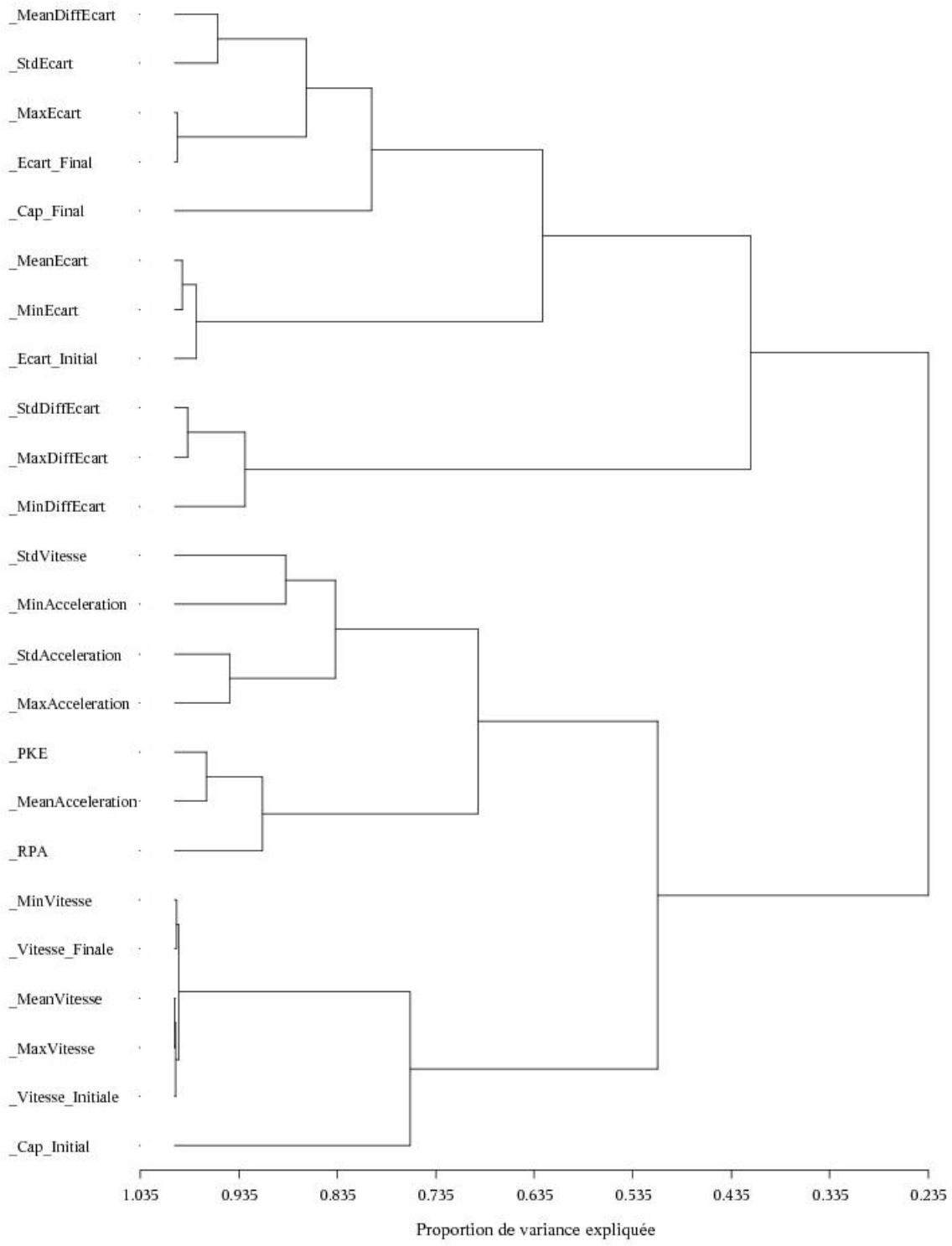


FIGURE 5.3 – *Arbre de la classification des indicateurs*

Classe	Variable	R^2 avec propre classe	R^2 avec le plus proche	Ratio 1-R^2
Cluster 1	**MeanVitesse**	**0.9964**	**0.0525**	**0.0038**
	MaxVitesse	0.9893	0.0344	0.0111
	MinVitesse	0.9869	0.0947	0.0145
	Vitesse Initiale	0.9764	0.0372	0.0245
	Vitesse Finale	0.9700	0.0657	0.0321
	Cap Initial	0.0573	0.0577	1.0000

TABLE 5.2 – *Cluster 1*

La première classe (table 5.2) rassemble tous les indicateurs liés à la *vitesse*, hormis son écart type : celui-ci étant naturellement lié aux indicateurs de l'*accélération*. Nous y trouvons également le *Cap Initial*, qui est si peu lié à la classe que nous ne nous étonnerons pas de sa présence ici. Cette classe semble traduire l'intervalle que recouvrent les *vitesses* ; la *vitesse* moyenne en est la plus représentative

Classe	Variable	R^2 avec propre classe	R^2 avec le plus proche	Ratio 1-R^2
Cluster 2	MeanEcart	0.9106	0.3559	0.1388
	MinEcart	**0.9465**	**0.1685**	**0.0643**
	Ecart Initial	0.8577	0.0487	0.1496

TABLE 5.3 – *Cluster 2*

La deuxième classe (table 5.3) rassemble les indicateurs liés à l'*écart*. *MinEcart* présente les meilleures caractéristiques (maximum de la première colonne et minimum des deux suivantes) et sera donc conservé.

Classe	Variable	R^2 avec propre classe	R^2 avec le plus proche	Ratio 1-R^2
Cluster 3	**StdDiffEcart**	**0.8943**	**0.1151**	**0.1195**
	MaxDiffEcart	0.8810	0.0683	0.1277
	MinDiffEcart	0.7308	0.1063	0.3012

TABLE 5.4 – *Cluster 3*

La troisième classe (table 5.4) rassemble les indicateurs liés à l'évolution des *écarts*. Nous choisirons ici de conserver *StdDiffEcart*, l'écart type des différences entre *écarts* successifs.

En prenant comme référence le marquage central, cet indicateur représente les changements de directions d'une trajectoire.

Classe	Variable	R^2 avec propre classe	R^2 avec le plus proche	Ratio 1-R^2
Cluster 4	**PKE**	**0.8182**	**0.1140**	**0.2052**
	RPA	0.7046	0.0715	0.3182
	MeanAcceleration	0.7937	0.0070	0.2077

TABLE 5.5 – *Cluster 4*

La quatrième classe (table 5.5) regroupe les variables liées à l'*accélération* générale. Le *PKE* présente les meilleures caractéristiques, et est de surcroît un indicateur communément utilisé, nous le conserverons pour représenter cette classe.

Classe	Variable	R^2 avec propre classe	R^2 avec le plus proche	Ratio 1-R^2
Cluster 5	**MeanDiffEcart**	**0.8340**	**0.0233**	**0.1700**
	StdEcart	0.6270	0.0879	0.4090
	MaxEcart	0.8375	0.5064	0.3292
	Ecart Final	0.8422	0.4820	0.3046
	Cap Final	0.1708	0.0152	0.8420

TABLE 5.6 – *Cluster 5*

La cinquième classe (table 5.6) regroupe les variables liées aux variations dans l'*écart*. *MeanDiffEcart* nous semble représenter le mieux cette classe car elle indique quelle direction prend globalement cette variable, de plus, c'est l'indicateur le moins corrélé aux autres classes. C'est donc cet indicateur que nous conserverons.

Classe	Variable	R^2 avec propre classe	R^2 avec le plus proche	Ratio 1-R^2
Cluster 6	StdVitesse	0.4323	0.0804	0.6173
	StdAcceleration	**0.8936**	**0.0862**	**0.1164**
	MaxAcceleration	0.5060	0.2331	0.6442
	MinAcceleration	0.6048	0.0312	0.4079

TABLE 5.7 – *Cluster 6*

La sixième classe (table 5.7) présente l'amplitude des variations de la *vitesse*. L'écart type de l'accélération est le mieux noté pour la représenter.

Nous conservons pour la suite de l'analyse les variables sélectionnées dans chaque classe. Elles sont présentées dans le tableau 5.8.

Variable	Signification
MeanVitesse	Vitesse moyenne
PKE	Accélération globale
StdAcceleration	Variations de l'accélérations
MinEcart	Position la proche du marquage central
MeanDiffEcart	Tendance de l'écart, par rapport au marquage central
StdDiffEcart	Variations autour de cette tendance

TABLE 5.8 – *Variables conservées*

5.2.3 Analyse en composantes principales

5.2.3.1 Les axes retenus par l'ACP

Les méthodes de classification permettent de classer nos trajectoires selon les six variables sélectionnées (tableau 5.8). En revanche, il est difficile de se représenter les spécificités de chaque classe. En réalisant une ACP, nous projetons les trajectoires sur un espace en 2 ou 3 dimensions, espace qu'il est plus aisé d'observer.

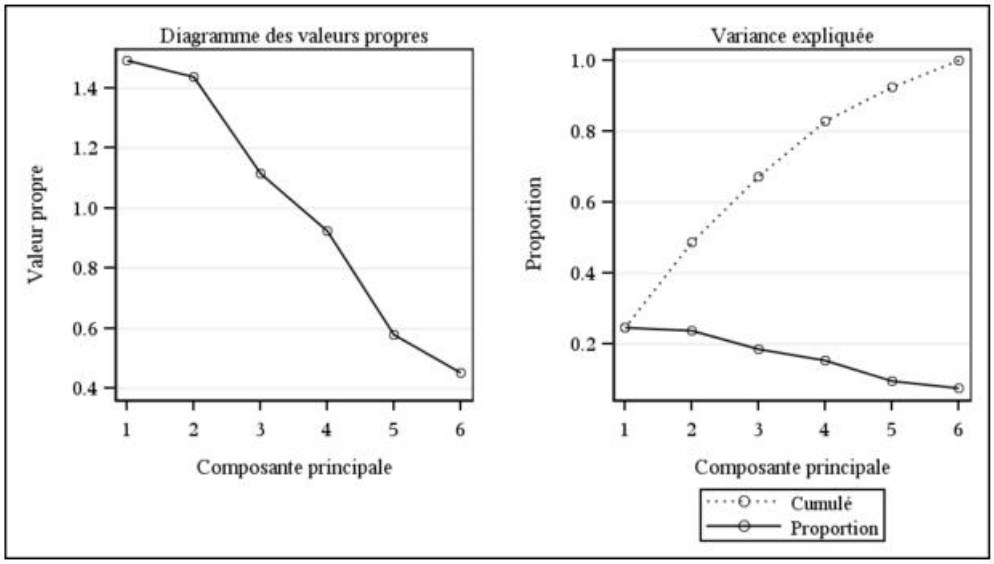


FIGURE 5.4 – *Résultat de l'ACP*

Le graphique 5.4 présente les composantes principales par leur valeur propre et la proportion de variance expliquée. Il existe 3 règles pour choisir le nombre de facteurs à retenir :

- La règle de Kaiser qui veut que l'on ne retienne que les facteurs aux valeurs propres supérieures à 1.
- On choisit le nombre de facteurs en fonction de la quantité d'information que l'on veut restituer.
- Le "scree-test" ou test du coude : on observe le graphique des valeurs propres et on ne retient que celles qui sont à gauche du point d'inflexion.

Le test du coude n'étant pas concluant, nous appliquerons la première règle et garderons donc les trois premières valeurs propres.

	Axe 1	Axe 2	Axe 3
MeanVitesse	0.2295	-0.3551	0.4449
PKE	0.0483	0.6128	0.2900
StdAcceleration	0.0415	0.6764	0.2081
MinEcart	-0.7118	0.0004	0.1632
MeanDiffEcart	-0.1523	-0.2018	0.7863
StdDiffEcart	0.6429	-0.0103	0.1728

TABLE 5.9 – *Vecteurs propres*

- Le premier axe est caractérisé par *MinEcart* et *StdDiffEcart*. Plus *MinEcart* y est faible, plus *StdDiffEcart* y est important, il différencie donc les véhicules qui s'approchent du marquage central avant de s'en écarter, de ceux qui restent à bonne distance du marquage central.
- Le deuxième axe est caractérisé par *PKE* et *StdAcceleration*. Il caractérise les véhicules qui ont une accélération forte et saccadée.
- Le troisième axe utilise les deux indicateurs restant : *MeanVitesse* et *MeanDiffEcart*. En mélangeant une variable spatiale et un indicateur lié à la vitesse, il décrit l'allure générale de la trajectoire : la direction et la vitesse moyenne.

Classons à présent les trajectoires selon leurs projections sur ces trois axes, avec la procédure VARCLUS de SAS décrite dans la section 5.2.2.1.

5.2.3.2 Classification descendante hiérarchique des trajectoires

Une partition en trois classes semble se dessiner, nous coupons l'arbre à la hauteur 12, et étudions chacune des trois classes.

Observons la représentation de ces classes de la figure 5.6.

- La première classe est la plus fournie : elle comporte 738 trajectoires, soit 65% de la base de données. Ses véhicules sont un peu plus lents que ceux des autres classes, et roulent à vitesse à peu près constante. Ils restent assez loin du marquage central et tournent assez peu.
- La deuxième classe a 385 trajectoires, soit 34% de la base de données. Elle contient les véhicules caractérisés par une accélération significative, une vitesse moyenne relativement importante, et une trajectoire tendant progressivement vers l'extérieur. De

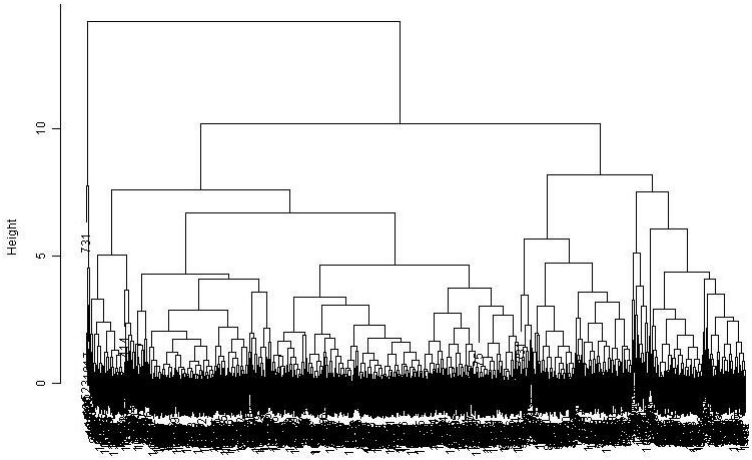


FIGURE 5.5 – *Dendogramme des trajectoires selon les 3 premières composantes principales*

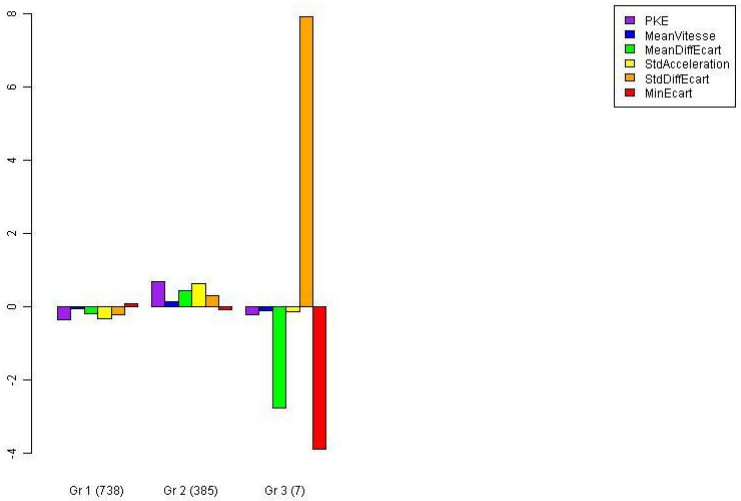


FIGURE 5.6 – *Histogramme des caractéristiques de chaque classe.*

nombreux sites relatifs à la course automobile recommandent de prendre un virage de l'extérieur, de passer par l'intérieur au coeur de ce virage pour terminer à l'extérieur. Cela permet de sortir du virage plus rapidement en accélérant dès la deuxième portion. Ces trajectoires semblent donc empruntées par des conducteurs expérimentés.

– La troisième classe comporte les 7 trajectoires exotiques. Elles sont caractérisées par leur position par rapport aux centre de la voie, et non par leur vitesse. Les véhicules de cette classe passent très près du marquage central (il s'avère qu'elles le traversent toutes), en outre, ils ont tendance à être loin du marquage central au départ, puis s'en rapprochent. Ces trajectoires sont dangereuses.

Ces classes ne semblent pas refléter les trois types de conduites présentés dans la section 2.1.3.

5.3 Un indice de risque fondé sur l'écart

5.3.1 Création de l'indice

Le risque que présente une trajectoire peut être déterminé par sa position sur la voie : si presque toutes les trajectoires passent par un même intervalle, c'est que celui-ci est plus sûr. Or pour déterminer ces intervalles, il nous faut utiliser un référentiel commun à toutes les trajectoires. La distance parcourue n'est pas la même pour tous les véhicules : certains zigzaguent et parcourent une plus grande distance que les véhicules suivant une trajectoire plus régulière. Nous aurions pu choisir la latitude, mais celle-ci n'augmente pas de façon linéaire par rapport à la distance parcourue : en effet vers la fin du virage, les incréments de la *latitude* sont inférieurs à ces mêmes incréments au début du virage. La *latitude* n'est donc pas non plus un bon choix de référentiel.

En revanche ce qui est constant pour toutes les trajectoires, c'est le marquage central. Nous prendrons donc comme référence l'abscisse curviligne du marquage central précédemment interpolé. Cette abscisse curviligne sera discrétisée en 64 positions. Il nous faut alors trouver les observations correspondant à ces positions.

Pour ce faire, nous interpolons les trajectoires, afin d'obtenir un large choix d'observations pour chaque position. Nous garderons l'observation la plus proche en termes de *latitude / longitude* de chaque position du marquage central. Nous obtenons bien 64 observations pour chaque trajectoire.

A partir de l'écart au centre de la voie, en chacune des positions, nous pouvons définir un intervalle regroupant une grande partie des trajectoires, la figure 5.7 présente la répartition des *écarts* au marquage central pour la première position. Nous observons que cette répartition est symétrique et qu'elle s'avère normale (le test de Kolmogorov Smirnov renvoie un p-value inférieure à 10^{-6}). On choisit alors d'exclure, par exemple, les 5% des plus petits *écarts* et les 5% des plus grands *écarts*, ce qui nous laisse un intervalle de confiance à 90%.

Nous calculons cet intervalle de confiance pour chacune des 64 positions. En reliant les bornes de ces 64 intervalles, nous obtenons un "faisceau de confiance" représentant un ensemble de trajectoires "admissibles" au sens statistique. Notre indice de risque d'une trajectoire est une variable binaire égale à :

- 0 si les *écarts* de notre trajectoire sont contenus dans le faisceau de confiance en chacune des 64 positions.
- 1 si notre trajectoire présente au moins un point en dehors de notre faisceau de confiance. Nous la considérerons alors comme "marginale".

Nous observons sur la figure 5.9 que, pour notre faisceau de confiance à 90%, 22% des trajectoires sont marginales.

5.3.2 Influence des variables liées à la vitesse sur la position sur la voie

Il est évident que les indicateurs liés à l'*écart* déterminent la *marginalité* d'une trajectoire. Nous cherchons maintenant à trouver un lien entre les indicateurs liés à la *vitesse* et la *marginalité* d'une trajectoire.

L'arbre de segmentation se lit ainsi :

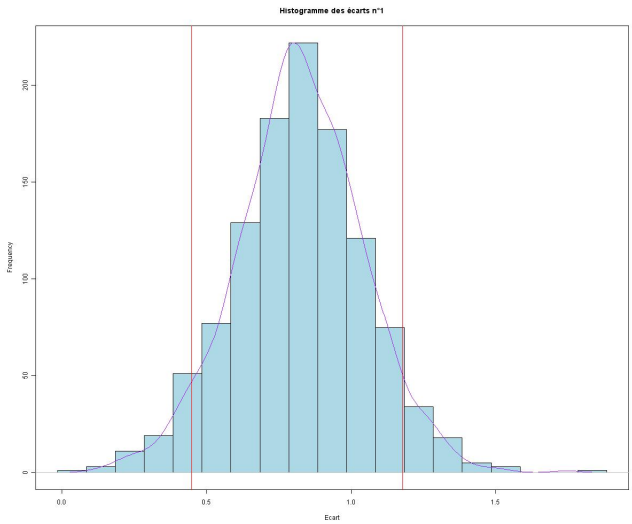


FIGURE 5.7 – *Histogramme des écarts en position 1. En rouge les limites de l'intervalle de confiance à 90%.*

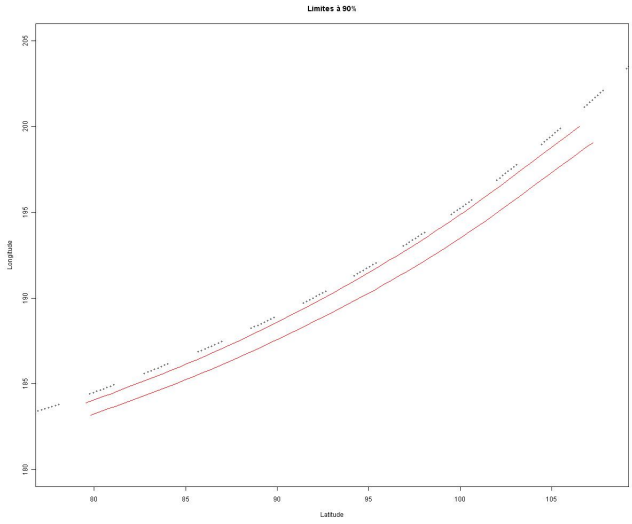


FIGURE 5.8 – *Faisceau de confiance des trajectoires.*

- 22.57% des trajectoires de l'échantillon initial sont marginales.
- Les trajectoires qui ont la plus forte accélération maximale (supérieure à $0.268m/s^2$) sont marginales (il y en a 5).
- Parmi les autres, 22% sont marginales.
- Si nous partageons ce même groupe entre les très lentes ($MaxVitesse < 47.7km/h$), et les autres, nous trouvons que 50% des plus lentes sont marginales (il y en a 20)
- Parmi les autres, 21% des trajectoires sont marginales.
- Si nous partageons ce dernier groupe entre les très rapides, et les normales, nous

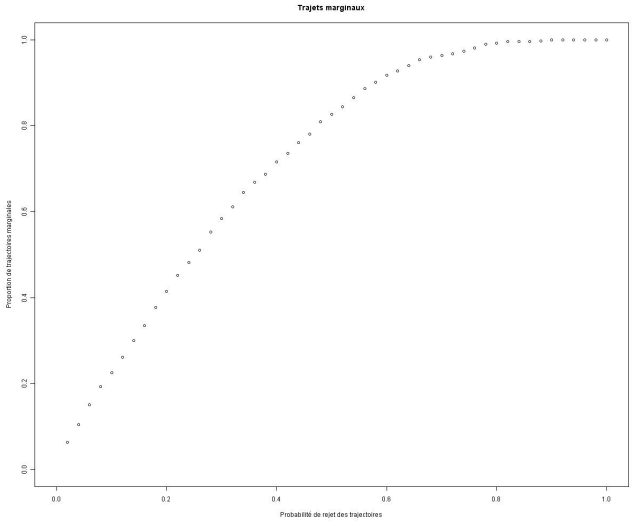


FIGURE 5.9 – *Proportion de trajectoires ayant au moins une observation en dehors de l'intervalle de confiance, en fonction de la proportion de données exclues de l'intervalle.*

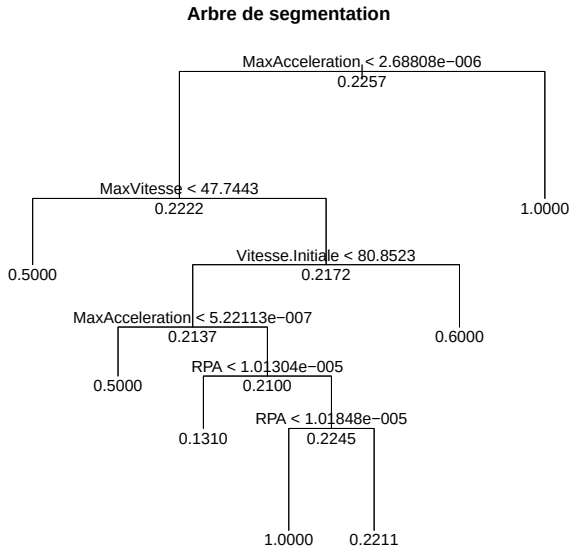


FIGURE 5.10 – *Arbre de segmentation des indicateurs liés à la vitesse pour discriminer les trajectoires présentant au moins une observation en dehors de l'intervalle des autres.*

trouvons que 60% des plus rapides sont marginales.

– Nous considérons à présent les trajectoires dont l'accélération n'est pas trop importante, et dont la vitesse est moyenne. C'est à présent les trajectoires qui ont une

accélération maximale faible qui sont pour la plupart marginales (à 50%), puis celles qui ont un RPA moyen (supérieur à $1.013.10^{-5}$ et inférieur à $1.0185.10^{-5}$) qui sont toutes marginales.

L'arbre de segmentation nous apprend que les trajectoires qui ont une *accélération maximale* forte sont toutes *marginales*, d'autres part, parmi les véhicules qui n'ont pas une *accélération maximale* forte, les véhicules lents sont pour moitié classés comme *marginaux*, mais globalement, trop peu de trajectoires expliquent ces noeuds. L'arbre de segmentation montre surtout qu'on ne peut détacher de grands groupes dont la proportion de trajectoires marginales est significative (soit très petites, soit très grande).

Il semble toutefois que la *vitesse* et l'*accélération* peuvent partiellement expliquer la position sur la voie.

Avec ces mêmes indicateurs, effectuons une régression logistique.

5.3.3 Régression logistique

La régression logistique étudie l'impact que peuvent avoir des variables numériques explicatives sur une variable dépendante binaire, c'est-à-dire qu'elle ne prend que deux valeurs (ici, $Y = 1$ si la trajectoire est marginale et $Y = 0$ sinon). Y suit alors une loi de Bernoulli caractérisée par les probabilités suivantes : $p = P(Y = 1)$ et $1 - p = P(Y = 0)$. L'objectif de la régression logistique est celui de toute régression : modéliser l'espérance conditionnelle $E(Y|X = x)$. On veut connaître la valeur moyenne de Y pour toute valeur de X. Pour une valeur Y valant 0 ou 1, cette valeur moyenne est la probabilité que $Y = 1$. On a donc : $E(Y|X = x) = P(Y = 1|X = x)$. Notons $\pi(x) = P(Y = 1|X = x)$ la probabilité que Y prenne la valeur 1 pour X fixé à x. La variable Y suit alors une loi de Bernoulli de paramètre $\pi(x)$. En régression logistique binaire simple, c'est-à-dire dans le cas d'une seule variable explicative X, on suppose que la probabilité $\pi(x)$ s'écrit sous la forme suivante :

$$\pi(x) = \frac{\exp(\beta_0+\beta_1 x)}{1+\exp(\beta_0+\beta_1)}.$$

De manière équivalente, on suppose que le rapport des côtes (ou odds ratio), aussi appelé logit, respecte la relation linéaire suivante :

$$log(\frac{\pi(x)}{1-\pi(x)}) = \beta_0 + \beta_1 x$$

Notons que la régression logistique binaire est en fait un cas particulier du modèle linéaire généralisé. Pour une variable Y suivant une loi de Bernoulli de paramètre $\pi(x)$, ce modèle s'écrit : $g(\pi(x)) = \beta_0 + \beta_1 x$ où la fonction g est appelée fonction de lien. En régression logistique, g est la fonction logit définie par g(x)=$\frac{x}{1-x}$ qui est l'inverse de la fonction logistique $F(x) = \frac{e^x}{1-e^x}$. La fonction logistique est bien adaptée à la modélisation de probabilités car elle prend ses valeurs entre 0 et 1 selon une courbe en S. Le modèle peut être étendu à l'analyse d'une variable réponse binaire Y en fonction de plusieurs variables explicatives $X_1, ..., X_k$ qui peuvent être quantitatives ou qualitatives. Le modèle logistique a alors pour expression :

$$\pi(x) = \frac{\exp(\beta_0+\beta_1 x_1+...+\beta_k x_k)}{1+\exp(\beta_0+\beta_1 x_1+...+\beta_k x_k)}$$

ou de manière équivalente :

$$log(\frac{\pi(x)}{1-\pi(x)}) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k$$

L'estimation du modèle de régression logistique consiste alors à estimer les coefficients β_i de ce modèle. Pour cela, on utilise la méthode du maximum de vraisemblance. En supposant que les observations sont indépendantes, cette vraisemblance, qui correspond à la probabilité d'observer les n données (x_i, y_i) calculée en fonction des paramètres inconnus β_i , s'écrit :

$$L(\beta_0, \beta_1, ..., \beta k) = \Pi_{i-1}^{n} \pi(x_i)^{y_i} (1 - \pi(x_i))^{y_i}$$

La maximisation de cette vraisemblance n'a pas de solution analytique, on utilisera donc des méthodes numériques afin d'obtenir les estimations $(\beta'_0, \beta'_1, ..., \beta'_k)$ des paramètres.

Etape	Effet Saisi	Effet Supprimé	DDL	Nombre dans	Khi-2 du score	Khi-2 de Wald	Pr > Khi-2
1	_StdAcceleration		1	1	4.8623		0.0275
2	_Vitesse_Finale		1	2	1.4553		0.2277
3		_Vitesse_Finale	1	1		1.4529	0.2281

FIGURE 5.11 – *Récapitulatif sur la sélection séquentielle de la procédure logistique.*

L'option STEPWISE de la procédure logistic sous sas suppose que le modèle entre d'abord une variable qui explique le mieux la variable dépendante, puis elle cherche la variable qui explique le mieux la variance qui reste à expliquer, ... et ainsi de suite en testant à chaque étape s'il n'est pas possible d'éliminer une variable entrée auparavant. Elle retient ainsi en priorité la variable *StdAcceleration*, avec une p-value à 2%. *StdAcceleration* est donc significative, les autres indicateurs, en revanche, ne le sont pas. Le second indicateur, supprimé par la regression logistique était ainsi la *vitesse finale*, avec une p-value à 20%, comme le montre la figure 5.11.

C'est ici l'indicateur StdAccel qui est retenu, alors que l'arbre décisionnel employait en premier lieu MaxAccel. Nous ne savons pas expliquer cette différence. En revanche, la marginalité semble, au vu de ces deux résultats liée à l'accélération plus qu'à la vitesse.

La courbe Roc (figure 5.12) nous apprend que, bien que significative, les variations de l'*accélération* ne suffisent pas à expliquer la *marginalité* d'une trajectoire.

5.3.4 Conclusions sur l'indice de risque

Nous n'avons pas pu démontrer la pertinence de notre indice de risque ni par un arbre de segmentation, ni par une regression logistique liant ce risque aux indicateurs du tableau 5.1. En effet, la trajectoire d'un véhicule est le résultat de phénomènes complexes liés à la vitesse et à la position de ce véhicule. D'une part, notre indice de risque ne prend en compte que la position de ce véhicule, et d'autre part nos indicateurs ne décrivent pas tous les aspects de la trajectoire. En outre, nous ne pouvons pas affirmer que l'échantillon contient bien des trajectoires à risque.

Il est donc difficile de valider ici notre indice de risque, mais nous ne pouvons pas non plus le rejeter de façon catégorique.

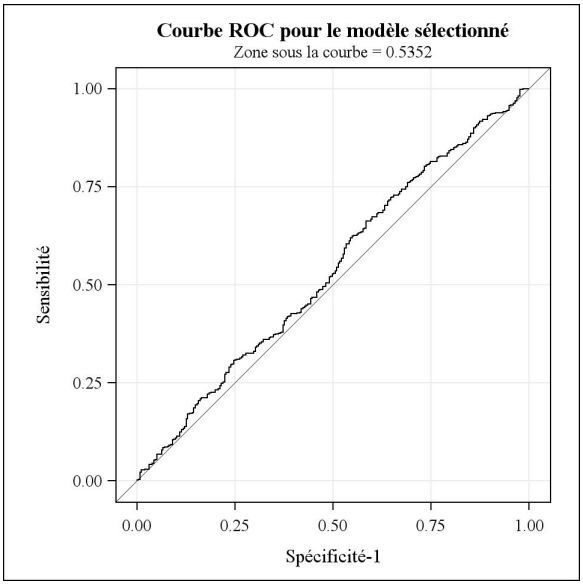


FIGURE 5.12 – *Courbe Roc de la regression logistique de l'indice de risque en fonction des indicateurs liés à la vitesse.*

Chapitre 6

Traminer

TraMineR est un package R proposé par Alexis Gabadinho, Matthias Studer, Nicolas S. Muller et Gilbert Ritschard et présenté lors d'une conférence à l'institut national d'études démographiques (INED) sur l'analyse de données longitudinales à laquelle nous avons assisté en avril 2010.

6.1 La méthode

TraMineR est un package R destiné à la description, l'exploration et la visualisation de séries d'états ou d'événements, ou plus généralement de séries de données discrètes. Son nom est la contraction de "Trajectory Miner for R". Ainsi il comprend des fonctions de visualisation de données et de leur typologie. Associé à une méthode de classification hiérarchique ascendante, il propose d'appliquer ces mêmes fonctions à chacune des classes.

6.1.1 La Classification Hiérarchique Ascendante (CAH)

La définition est tirée de Tufféry [2007] : Tout comme la classification hiérarchique descendante, vue au chapitre précédent, la CAH produit des suites de partitions emboîtées. Celles-ci sont d'hétérogénéités croissantes entre la partition en n classes où chaque objet est isolé, et la partition en une classe qui regroupe tous les objets. La CAH est utilisable dès que l'on dispose d'une notion de distance que ce soit dans un espace des individus ou dans un espace des variables. Il faut avoir défini la distance entre deux objets, qui est généralement naturelle, et la distance entre deux classes, qui laisse plus de possibilités. La méthode de Ward est l'une de celles qui correspondent le mieux à l'objectif de la classification. Comme une bonne classification est une classification pour laquelle l'inertie interclasse est élevée, et comme le passage d'une classification en $k+1$ classes à une classification en k classes (regroupement de 2 classes) ne peut que faire baisser l'inertie interclasse, on cherchera à fusionner les deux classes qui feront le moins baisser l'inertie interclasse. La notion de distance correspondant à cet objectif est la distance de Ward entre deux classes définie ci-dessous :

$$d(A,B) = \frac{d(a,b)^2}{\frac{1}{n_A}+\frac{1}{n_B}},$$

avec A et B les deux classes, de barycentres a et b et d'effectifs n_A et n_B.

La méthode de Ward est de loin la méthode la plus utilisée en classification ascendante hiérarchique car elle s'applique bien aux problèmes réels.

6.1.2 L'appariement optimal

La méthode d'appariement optimal s'appuie sur un ensemble d'algorithmes dynamiques utilisés principalement par la biologie moléculaire pour analyser les similarités entre chaînes d'ADN. Elle a ensuite été introduite dans les sciences sociales par Andrew Abbott dans les années 1980. Son principe est basé sur la notion de similarité ou de dissimilarité entre des paires de séquences. L'idée de base consiste à mesurer la dissimilarité entre deux séquences en évaluant le coût représenté par la transformation de l'une des séquences en l'autre. La transformation est effectuée au moyen de 3 opérations élémentaires :

- L'insertion : un élément est inséré dans la séquence.
- La suppression : un élément est supprimé dans la séquence.
- La substitution : Un élément est substitué à un autre.

On peut assigner un coût spécifique à chacune de ces opérations élémentaires. Une série d'opérations a un coût équivalent à la somme des coûts des opérations élémentaires. La distance entre deux séquences est alors définie comme le coût minimal de la transformation d'une séquence en l'autre. L'appariement de l'ensemble des paires de séquences aboutit à la création d'une matrice de distance, que l'on peut ensuite utiliser pour regrouper les séquences les plus similaires.

6.2 Application à l'accélération transversale

TraMineR analyse une unique variable, or l'ACP nous a montrée que les trajectoires se différencient non seulement par leur position, mais aussi par leur vitesse. Aucune des variables retenues (tableau 4.4) ne décrit ces deux notions, mais l'*accélération latérale* décrit la relation entre vitesse et courbe. C'est donc cette variable que nous analyserons.

Elle est fondée sur le *rayon de courbure*, lequel est très sensible aux erreurs de mesure. Ses résultats peuvent, malgré les lissages effectués, se révéler approximatifs.

6.2.1 Discrétisation de l'accélération latérale

TraMiner analyse une variable à travers ses états. Il nous faut donc discrétiser la variable *accélération latérale*. Nous observons sur la figure 6.1 que l'*accélération latérale* semble suivre une répartition normale, donc symétrique. Nous aurons donc un nombre impair de classes, afin d'observer la classe centrale. Nous choisissons d'utiliser cinq classes, car de trop nombreuses classes seraient difficiles à visualiser. Les quantiles représentés sur cette même figure nous offrent des points de discrétisation permettant d'obtenir un même nombre de valeurs dans chaque classe.

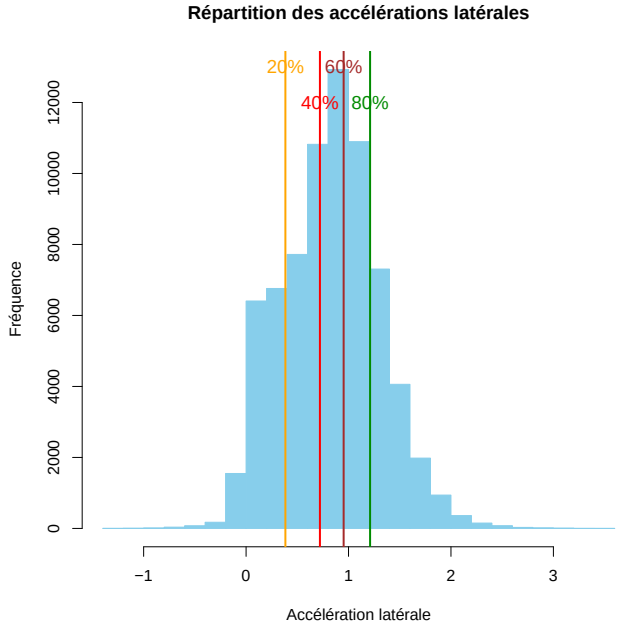


FIGURE 6.1 – *Répartition des accélérations latérales en $m.s^{-2}$. Les barres verticales représentent les quantiles à 20% (0.38 $m.s^{-2}$), 40% (0.72 $m.s^{-2}$), 60% (0.95 $m.s^{-2}$), et 80% (1.21 $m.s^{-2}$)*

6.2.2 Classification

L'arbre hiérarchique (figure 6.2) présente deux grandes classes qui se divisent elles mêmes en deux classes presque au même niveau. Afin de retrouver la classification en trois types de conduites, nous étudierons ici la classification en trois classes.

TraMineR permet de visualiser ces classes selon deux aspects : l'évolution de toutes les trajectoires de la classe, et l'évolution de la répartition de ces trajectoires.

Nous observons que toutes les trajectoires semblent commencer en ligne droite, ce qui doit être du à l'extrapolation des données, puis au lissage de celles-ci. En effet nous avons du appliquer un spline nous fournissant la dérivée des données, mais pour la première donnée, la direction n'étant pas encore définie, l'*accélération latérale* est nulle. Cette partie n'est donc pas à considérer pour l'*accélération latérale*, mais la zone d'étude des données étant déjà très restreinte, nous avons préféré la garder.

Nous observons sur les figures 6.3 et 6.4 que :

- La première classe (481 trajectoires) présente deux phases où l'accélération latérale est forte séparées par une phase où l'accélération latérale est plus faible.
- La deuxième classe (528 trajectoires) a la même allure que la première classe, mais avec une *accélération latérale* partout plus forte.
- La troisième classe (121 trajectoires) diffère des deux précédentes, car elle ne présente qu'un virage, pris assez tôt, avec une forte *accélération latérale*.

Observons à présent les indicateurs sélectionnés plus tôt pour chaque classe ; ces variables ont été centrées et réduites afin d'être visualisées sur la figure 6.5.

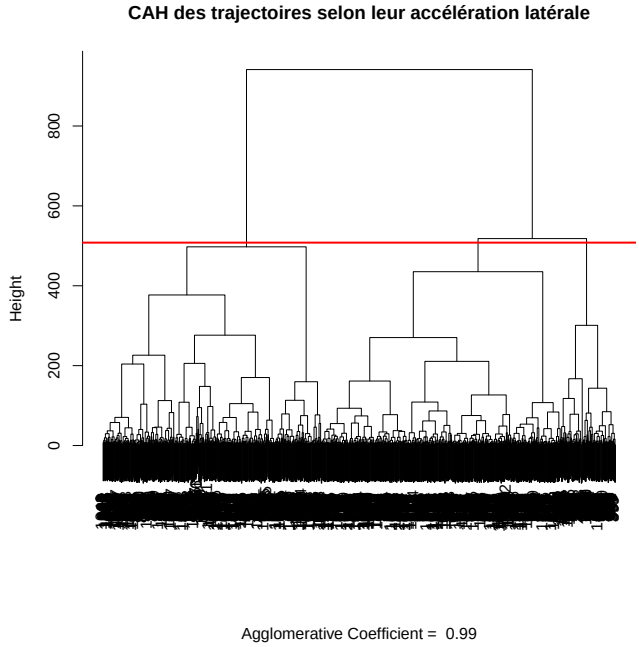


FIGURE 6.2 – *Arbre de la CAH sur les accélérations latérales*

On voit ici que les indicateurs différencient clairement les trajectoires :

- La première classe est composée de véhicules très éloignés du marquage central (*MinEcart* faible), qui dévient peu de leur trajectoire (*StdDiffEcart* faible). Leur vitesse (*MeanVitesse*)est assez faible, et ils ont tendance à accélérer (*PKE* important).
- Les véhicules de la deuxième classe sont les plus rapides (*MeanVitesse* important), mais sont aussi ceux qui accélèrent le moins (*PKE* très faible). Leur position sur la voie est moyenne (*MinEcart*), mais varie largement en se rapprochant du marquage central (*StdAcceleration* et *MeandDiffEcart* fort).
- Les véhicules de la troisième classe sont assez rapides et accélèrent encore (*MeanVitesse* et *PKE* importants). Ils passent très près du marquage central, et ont tendance à s'en écarter.

Il est difficile de discerner les trois styles de conduites proposés dans l'état de l'art : Notre première classe est ici constituée de véhicules lents qui restent éloignés du marquage central, la deuxième classe est constituée de véhicules très rapides qui freinent et se rapprochent du marquage central, ils prennent donc leur virage un peu tard. La troisième classe serait celle du troisième comportement décrit dans l'état de l'art : des véhicules rapides qui suivent une trajectoire "extérieur-intérieur-extérieur" afin de limiter la décélération.

L'*accélération latérale* semble donc bien être une variable capable de traduire les caractéristiques géométriques ainsi que les caractéristiques cinétiques d'une trajectoire.

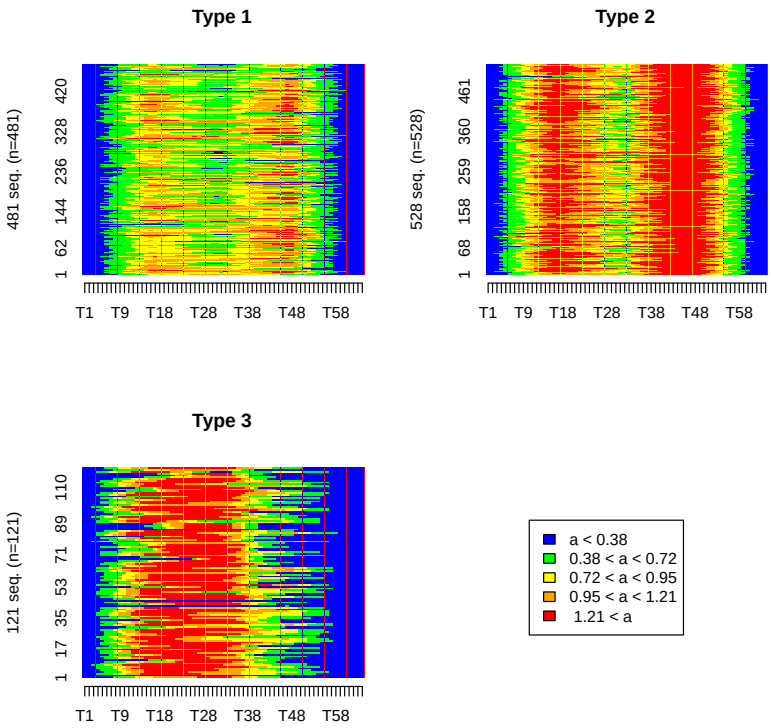


FIGURE 6.3 – *Evolution de l'accélération latérale dans les 3 classes*

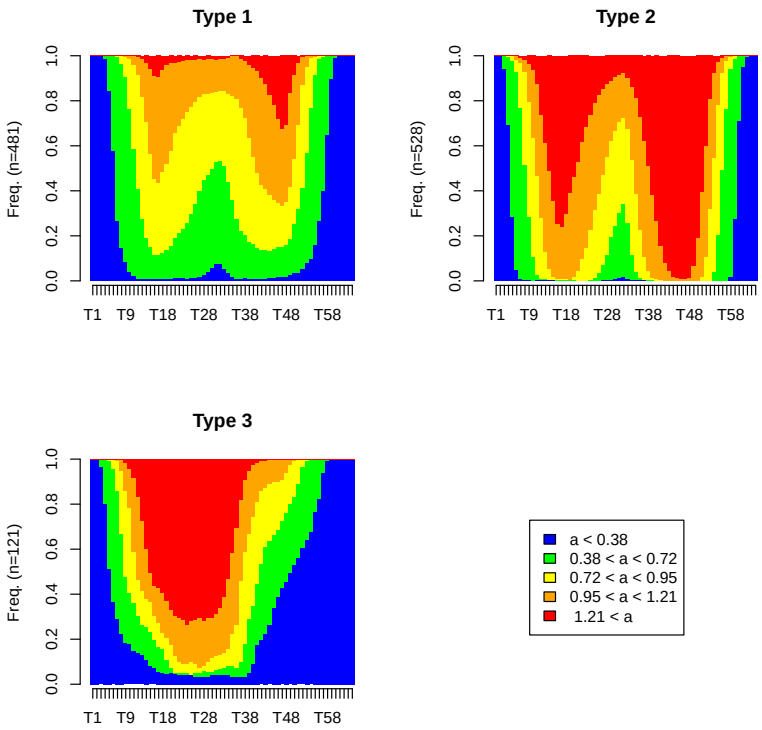


FIGURE 6.4 – *Evolution de l'accélérations latérales dans les 3 classes*

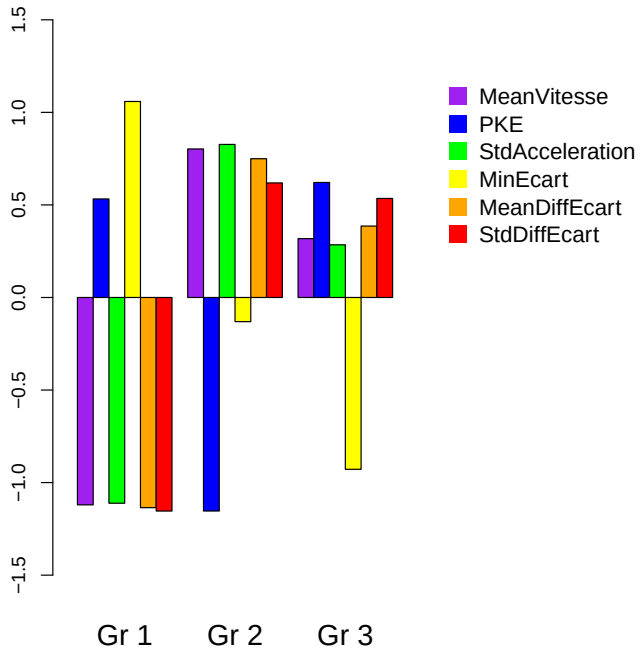


FIGURE 6.5 – *Moyennes des variables centrées réduites sélectionnées pour l'ACP, pour chaque groupe établi par TraMineR.*

Chapitre 7

Kml

Le package R KML est proposé par Christophe Genolini. Nous l'avons découvert en même temps que le package TraMineR lors d'une conférence à l'INED. Il est destiné à l'analyse de données longitudinales et à leur classification par la méthode des K-moyennes.

7.1 La méthode

7.1.1 La méthode des K moyennes

Méthode des K moyennes est une méthode de descente (c'est-à-dire qu'à chaque pas de recherche, cette méthode progresse vers une solution voisine de meilleure qualité) appartenant à la classe des algorithmes EM (Maximisation de l'espérance). Les algorithmes EM procèdent comme suit : au départ, chaque observation est assignée à un cluster. Viennent ensuite deux phases : une étape d'évaluation de l'espérance (E), où l'on calcule l'espérance de la vraisemblance en tenant compte des dernières variables observées, et une étape de maximisation (M), où l'on estime le maximum de vraisemblance des paramètres en maximisant la vraisemblance trouvée à l'étape E. On utilise ensuite les paramètres trouvés à l'étape M comme point de départ d'une nouvelle phase d'évaluation de l'espérance, et l'on itère ainsi.

7.1.2 Les critères d'arrêt

On trouve, dans la littérature statistique, peu d'articles traitant de la comparaison entre méthodes pour déterminer le nombre de classes d'une base de données en général. On en réfère donc souvent à la même étude, proposée par Glenn W. Milligan et Martha C. Cooper en 1985 (Milligan et Cooper [1985]). La méthode qu'ils ont employée prévoyait une base de données séparée en 2,3,4 ou 5 classes distinctes, les données appartenant à des espaces euclidiens à 4,6,8 dimensions. Les 30 critères d'arrêts sélectionnés provenaient d'une grande variété de sources et de disciplines. Cette étude aboutit à l'existence d'un critère optimal dans leur expérience, le critère de Calinski et Harabasz égal à :

$$\frac{Trace(B)}{k-1} \cdot \frac{n-k}{Trace(W)}$$

où n est le nombre total de trajectoires, k le nombre optimal de classes (il maximise ce critère), W la matrice de corrélations intra-classes, et B la matrice de corrélations inter-classes. C'est donc ce critère que nous utiliserons.

7.2 Application à l'*accélération latérale*

Comme au chapitre précédent, et pour les mêmes raisons qui y sont évoquées (cf section 6.2), c'est l'*accélération latérale* qui sera retenue pour comparer les trajectoires. En revanche, contrairement à TraMineR, le package KML peut comparer des variables quantitatives. Nous n'aurons donc pas à discrétiser l'*accélération latérale*.

7.2.1 Classification

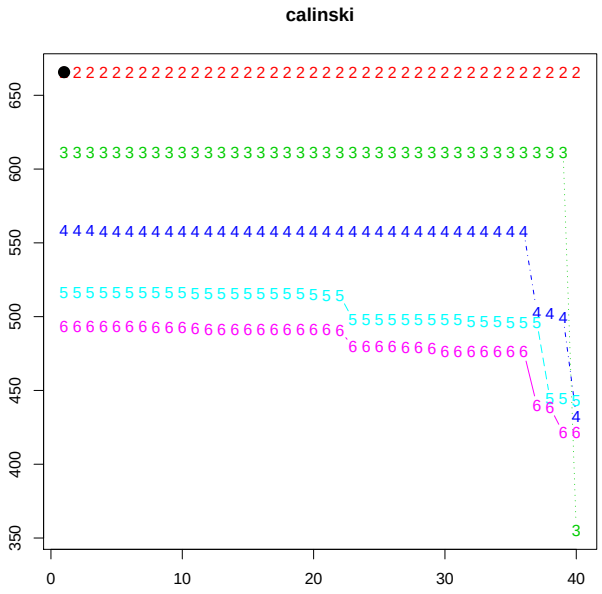


FIGURE 7.1 – *Critère de Calinski et Harabasz.*

La figure 7.1 présente le critère de calinski pour des nombres de classes différents et des conditions initiales différentes. En effet la méthode des K moyennes dépend également des conditions initiales. KML teste donc la méthode pour des conditions initiales différentes, et renvoie les meilleurs résultats. On observe que plus il y a de classes, plus le critère de Calinski est faible, ce qui n'est pas inhérent à ce critère. Cela peut signifier que les données ne sont pas suffisamment variées, et qu'il est difficile de les classer par cette méthode, ou encore que deux classes suffisent pour discerner ces données.

Comme dans le chapitre précédent, nous étudierons une classification en 3 clusters, afin de chercher les 3 types proposés dans l'état de l'art, classification en 3 clusters qui par ailleurs a un bon critère de Calinski.

Nous observons sur la figure 7.2 deux premières classes de même disposition, en M, différentes par l'amplitude de ses pics, et une troisième classe tout à fait différentes des deux premières.

Observons à présent les indicateurs sélectionnés plus tôt pour chaque classe ; ces variables ont été centrées et réduites afin d'être visualisées sur un même graphique (figure 7.3).

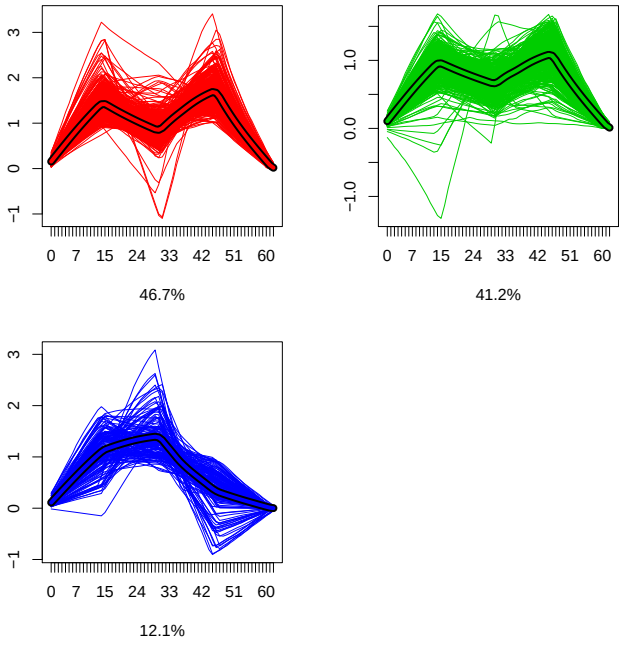


FIGURE 7.2 – *Visualisation de la classification en trois classes : classe 1 en rouge, classe 2 en vert, classe 3 en bleu.*

On voit ici que les indicateurs différencient clairement les trajectoires :

- La première classe est composée de 528 véhicules très lents (*MeanVitesse*), et éloignés du marquage central (*MinEcart*). Cet éloignement variant peu durant le parcours (*StdDiffEcart*).
- La deuxième classe est composée de 465 véhicules très rapides (*MeanVitesse*) et qui n'accélèrent pas du tout (*PKE* très faible), ceux-ci sont assez éloignés du marquage central (*MinEcart*), cet éloignement variant peu durant le parcours (*StdDiffEcart*).
- La troisième classe est composée de 137 véhicules qui passent très près du marquage central (*MinEcart*) et s'en écartent fortement (*MeanDiffEcart*). Leur vitesse est moyenne (*MeanVitesse*), mais leur accélération est très forte (*PKE*).

Ces trois classes sont très proches des classes trouvées précédemment avec le package TraMineR, on trouve d'ailleurs 90% de trajectoires classées de la même manière.

Lorsque l'on s'intéresse de plus près à la troisième classe, qui présente selon nous les caractéristiques du type de conduite n°3, on observe que 84% des trajectoires présentes dans le groupe 3 de TraMineR sont présentes dans le groupe 3 de KML, et que 75% des trajectoires du groupe 3 de KML sont présentes dans le groupe 3 de TraMineR.

En revanche, cette troisième classe disparaît lorsque l'on passe à la classification en deux classes, validée par le critère de Calinski (figure 7.4). On observe en effet ici que l'*accélération latérale* moyenne des deux classes est en forme de M.
L'hypothèse selon laquelle l'*accélération latérale* serait capable de traduire les caractéristiques géométriques ainsi que les caractéristiques cinétiques d'une trajectoire semble ici corroborée.

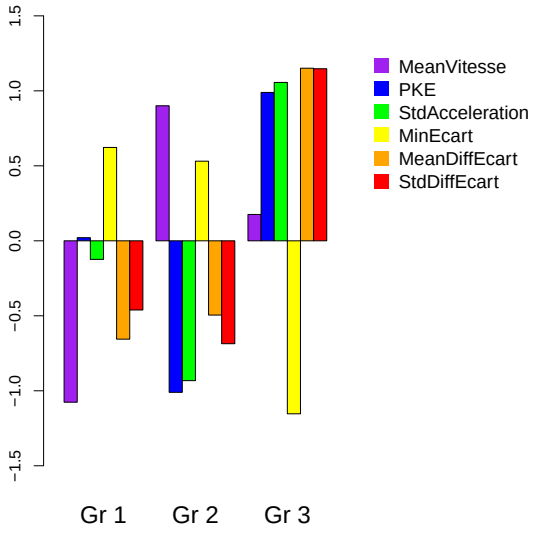


FIGURE 7.3 – *Moyennes des variables centrées réduites sélectionnées pour l'ACP, pour chaque groupe établi par KML.*

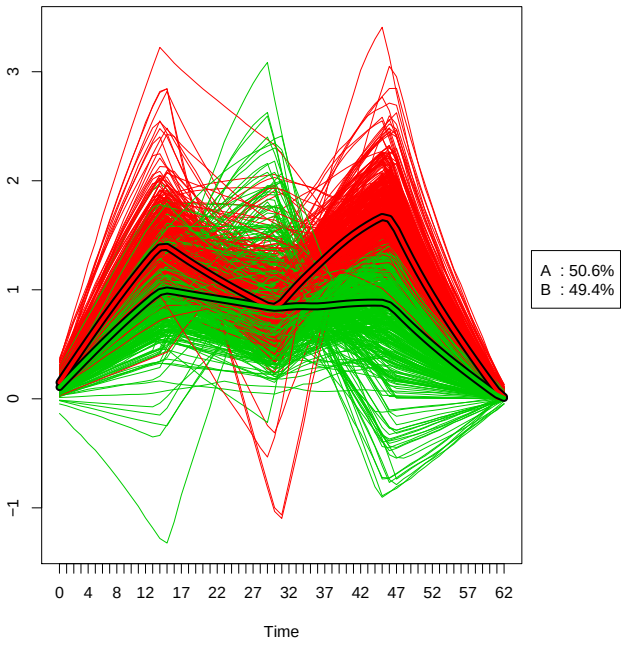


FIGURE 7.4 – *Visualisation de la classification en deux classes.*

Chapitre 8

Classification à partir de la Dynamic Time Warping

Nous avons déjà décrit cette distance dans la section 3.1.4. Son principal avantage est de pouvoir rapprocher des trajectoires identiques après translation, ou encore après dilatation. Elle sera ici appliquée à des données de même longueur, mesurées aux mêmes endroits. D'autres distances pourraient être utilisées, comme nous l'avons vu précédemment, mais il est intéressant de voir les résultats qu'elle fournit.

8.1 La méthode

La distance Dynamic Time Warping sera ici associée à la distance euclidienne pour calculer la distance entre les *accélérations latérales* de nos trajectoires. Comme la distance DTW ne renvoie pas le même résultat pour le calcul de la distance DTW entre "référence" et "requête" et le calcul de la distance DTW entre "requête" et "référence" (elle n'a pas la propriété de symétrie), nous prendrons le minimum des deux distances DTW obtenues pour chaque couple de trajectoires pour construire notre matrice de distances.

Nous appliquerons ensuite la méthode de Ward (cf section 6.1.1) à cette matrice de distances.

8.2 La classification

Ici encore, pour chercher dans nos données les trois types de comportement évoqués dans la section 2.1.3, et pour comparer notre méthode aux précédentes classifications, nous classerons nos trajectoires en trois clusters. L'arbre hiérarchique (figure 8.1) présente ici trois classes distinctes.

Observons à présent les indicateurs sélectionnés plus tôt pour chaque classe ; ces variables ont été centrées et réduites afin d'être visualisées sur un même graphique (figure 8.2).

On voit ici que les indicateurs différencient clairement les trajectoires :

– La première classe est composée de 616 véhicules lents (*MeanVitesse*) qui n'accélèrent pas (*PKE* et *StdAcceleration*). Ils restent loin du marquage central (*MinEcart*), bien qu'ils tendent à s'en approcher (*MeanDiffEcart*)

– La deuxième classe est composée de 343 véhicules rapides qui ont tendance à ralentir. Ils roulent au milieu de la voie (*MinEcart*), et ont tendance à y rester (*MeanDiffEcart* et *StdDiffEcart*).
– La troisième classe est composée de 171 véhicules qui passent très près du marquage central (*MinEcart*) et s'en écartent fortement (*MeanDiffEcart*). Leur vitesse est moyenne (*MeanVitesse*), mais leur accélération est très forte (*PKE*).

Si les deux premières classes ne sont assimilables ni à celles obtenues par les classifications précédentes, ni aux comportements de conduite recherchés, la troisième classe semble correspondre au troisième type de conduite.

Ainsi 80% des trajectoires sont classées de la même façon par KML et par la DTW, et 74% des trajectoires sont classées de la même façon par TraMiner et par la DTW. D'autres parts, 98% des trajectoires de la classe 3 obtenu par KML se trouvent dans la classe 3 obtenue par la dynamic time warping. De même 87% des trajectoires de la classe 3 obtenu par TraMineR se trouvent dans la classe 3 obtenue par la dynamic time warping.

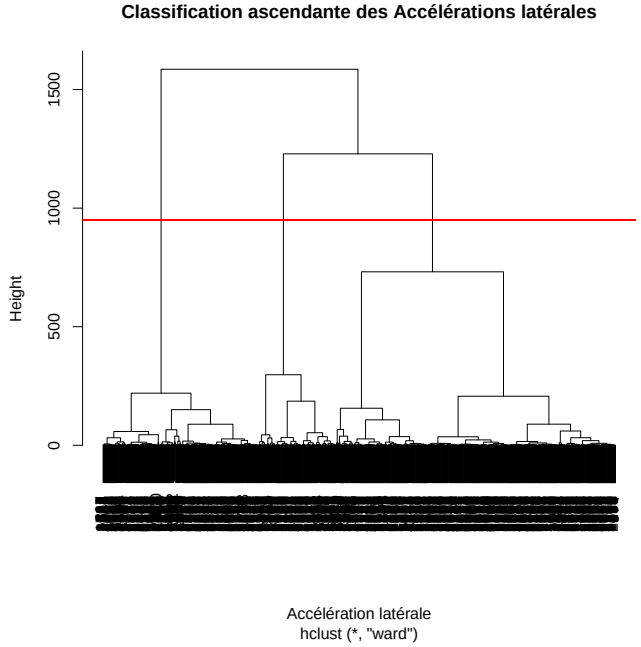


FIGURE 8.1 – *Arbre de la CAH sur les accélérations latérales, obtenue grâce aux dynamic time warping*

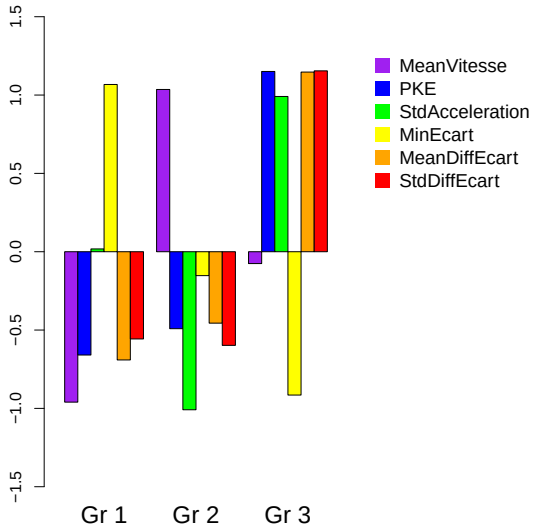


FIGURE 8.2 – *Moyennes des variables centrées réduites sélectionnées pour l'ACP, pour chaque groupe obtenu grâce à la dynamic time warping.*

Chapitre 9

Comparaison des méthodes

9.1 Traitement de chacune des méthodes

9.1.1 Analyse par indicateurs

La méthode d'analyse par indicateurs nécessite un calcul de très nombreux indicateurs, qui tendent à recouvrir l'ensemble des caractéristiques des trajectoires, lesquelles ne sont plus décrites par plusieurs observations. Puis il faut regrouper les indicateurs par classes afin d'effectuer une sélection ; cela allège encore les calculs suivants mais provoque une perte d'information, certes minimisée par les méthodes de classification hiérarchique. Il faut encore effectuer une ACP pour combiner les indicateurs restants en une nouvelle base, et nous pouvons enfin classer les trajectoires.

9.1.2 Analyse par TraMineR

TraMineR nécessite des données de même taille recouvrant la même zone. Il faut donc interpoler ces données. Puis vient ensuite le calcul de l'*accélération latérale*, variable décrivant la relation entre la géométrie de la trajectoire et sa dynamique. Cette variable étant très sensible, il nous faut d'abord lisser les données.

TraMineR étudiant des données qualitatives, il nous faut les discrétiser, ce qui provoque une perte d'information due à l'approximation et à la perte d'ordre hiérarchique entre les données. Enfin nous pouvons effectuer la classification par une méthode hiérarchique.

9.1.3 Classification par la méthode des K moyennes

Comme pour TraMineR, nous étudions avec KmL une seule variable : l'*accélération latérale* des trajectoire ; ce qui induit les traitements précités. En revanche KmL traite des données quantitatives, et la classification des observations vient juste après le calcul de la variable.

9.1.4 Classification à partir de la distance dynamic time warping

Nous avons ici aussi étudié l'accélération latérale des trajectoires, sous leur forme initiale. Il a fallu dans un premier temps calculer toutes les distances, deux fois pour chaque couple de trajectoires, ces distances n'étant pas symétriques. Nous avons ensuite retenu le minimum

de chaque couple de distances, puis effectué une classification hiérarchique ascendante, avec la méthode de Ward, à partir de la matrice des distances.

9.2 Comparaison des classes

On observe tout d'abord que la classification en deux classes par la méthode des K moyennes est très proche de la classification en deux classes par TraMineR : seulement 15% des trajectoires ne sont pas classées de la même façon. En revanche, si l'on compare ces mêmes méthodes pour une classification en six clusters, les classes sont tout à fait différentes, car si TraMineR utilise les partitions emboîtées d'une classification hiérarchique, les classes de la méthodes des K moyennes sont reconstruites entièrement quand on change le nombre de clusters.

Réduisons à présent notre comparaison aux classifications en trois classes, présentée dans le tableau 9.1 :

- Les classes obtenues à partir des indicateurs (notées ACP) sont très différentes des autres. Les valeurs importantes trouvées dans les colonnes 1 et 2 de l'ACP sont dues aux nombres importants de trajectoires qui composent ses classes 1 et 2.
- Le tableau n'est pas symétrique, car l'intersection de deux classes est normalisée par le nombre de trajectoire contenue par l'une d'entre elles. On trouve pourtant de fortes valeurs en (i, j) et en (j, i), pour i et j déterminant une des trois classes définies par l'une des trois méthodes : TraMineR, KML, et DTW. Les classes 1, 2 et 3 sont donc très semblables entre les méthodes TraMineR, KML et DTW.
- Il est à noter que les classes de la DTW sont tout de même assez différentes. On trouve en effet des valeurs inférieures à 80 lorsque l'on prend le minimum de (i, j) et (j, i), pour i une classe de la méthode DTW et j une classe de la méthode TraMineR ou KML.

La similitude entre les classes obtenues par les méthodes KML, TraMineR et DTW est due principalement au fait que ces trois méthodes ont été appliquées à la même variable *accélération latérale*. La première classification employait, quant à elle, des indicateurs liés à la vitesse et à la position des véhicules, et non l'*accélération latérale*. C'est pour cette raison que les classes obtenues par cette méthode sont bien différentes des classes obtenues par les trois autres méthodes.

Comparons à présent ces mêmes classes décrites à l'aide des indicateurs conservés pour l'ACP.

Concernant la classification par indicateurs, la figure 5.6 nous montrait deux première classes moyennes, opposées à une troisième classe très différente par les écarts de ses trajectoires. Elle mettait ainsi en évidence les 7 trajectoires exotiques de la classe 3.

Concernant la classification par TraMineR, la figure 6.5 nous montrait une première classe de trajectoires lentes et éloignées du marquage central, une deuxième classe de trajectoires rapides et assez proches du marquage central, et une troisième classe semblant correspondre au 3ème type de conduite présenté dans la section 2.1.3.

Concernant la classification par KML, la figure 7.3 nous montrait deux première classes différentes surtout par la vitesse de leurs trajectoires, la troisième ressemblant au 3ème type de conduite, avec un écart au marquage central très petit.

Concernant la classification par la DTW, la figure 8.2 nous montrait des classes assez semblables à la précédente, à celà près que les trajectoires de la classe 2 avait une un écart

au centre de la voie bien inférieur à celui des trajectoires de la classe 1.

Méthode		ACP			TRA			KML			DTW		
	Classe	1	2	3	1	2	3	1	2	3	1	2	3
ACP	1				39	39	22	36	39	26	43	27	30
	2				44	51	5	44	51	5	61	32	7
	3				43	43	14	43	43	14	43	43	14
TRA	1	31	68	1				88	5	7	85	1	14
	2	28	71	0				6	94	0	37	63	0
	3	71	28	1				11	5	85	10	9	87
KML	1	29	70	1	91	6	3				93	0	7
	2	28	71	1	5	95	1				34	65	1
	3	72	27	1	26	0	74				2	0	98
DTW	1	27	73	0	66	32	2	70	29	0			
	2	31	69	1	2	97	1	0	100	0			
	3	68	32	1	39	0	61	19	2	78			

TABLE 9.1 – *Pourcentage de trajectoires de la classe "ligne" appartenant aussi à la classe "colonne". Les lignes et les colonnes sont définies par les trois classes des méthodes : Classification à partir d'indicateurs (ACP), par les fonctions de TraMineR (TRA), par les fonctions de KML (KML) ou par la classification obtenue à partir de la distance Dynamic Time Warping (DTW).*

9.3 L'indice de risque

En ce qui concerne l'indice de risque établi dans le chapitre 5.3, seule la troisième classe semblable dans les classifications présente une proportion de trajectoires *marginales* significative, les autres classes en étant trop peu pourvues. La figure 9.1 montre que quelque soit la méthode utilisée, environ un tiers des trajectoires de la troisième classe sont *marginales*, tandis que moins d'un quart de l'ensemble des trajectoires sont marginales. Or cette troisième classe correspondrait aux véhicules suivant la conduite de type 3, conduite employée par des conducteurs expérimentés. Cela peut remettre en question notre indice de risque, qui n'est fondé que sur la position du véhicule par rapport à la position de l'ensemble des véhicules. D'autre part, la troisième classe de la méthode de classification par indicateurs est composée intégralement de trajectoires marginales.

Il existe donc un vrai lien le troisième type de conduite et l'indice de risque.

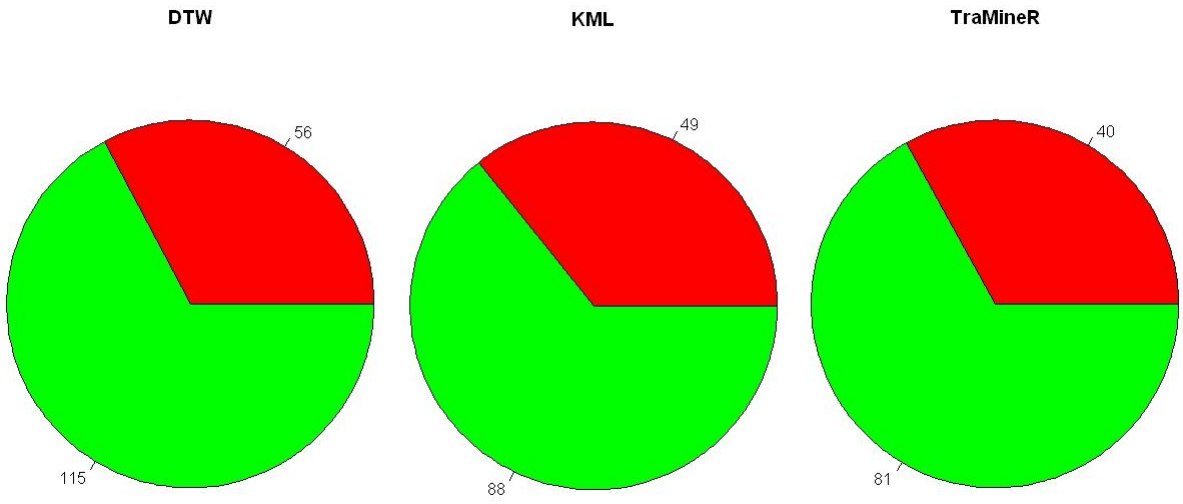


FIGURE 9.1 – *Pourcentage de trajectoires marginales (en rouge) et pourcentage de trajectoires "normales" (en bleu) pour la classe 3 de Dtw, KML et TraMineR de gauche à droite.*

9.4 Avantages et inconvénients

- La méthode de classification à partir des indicateurs est assez longue établir, mais une fois les indicateurs sélectionnés, on peut les réutiliser dans de prochaines analyses. Ces indicateurs permettent de mettre en évidence les trajectoires particulières, qui tendent à zigzaguer. La méthode ne prévoit pas de visualisation des trajectoires.
- La méthode de classification par TraMineR est assez longue, car elle nécessite le calcul de l'*accélération latérale*, puis sa discrétisation. La classification elle-même ne prenant pas trop de temps. Elle met en évidence une troisième classe qui se rapproche du troisième type de conduite, et offre des méthodes de visualisation claires et agréables.
- La méthode de classification par KML est très rapide, une fois que l'*accélération latérale* a été calculée. Elle met en évidence une troisième classe qui se rapproche du troisième type de conduite, mais propose un critère qui arrête la classification à deux classes. Cette méthode prévoit la visualisation de chaque classe.
- La méthode de classification à partir de la distance DTW est très lente (de l'ordre de quelques heures), car elle nécessite le calcul des distances DTW entre toutes les trajectoires. Son résultat est assez semblable aux deux précédents, puisqu'elle met elle aussi en avant une troisième classe semblable au troisième type de conduite.

Pour être validées, ces méthodes doivent être testées sur des trajectoires plus longues et plus nombreuses, avec l'appui de connaissances des relations entre trajectoire et état d'esprit du conducteur.

Chapitre 10

Conclusion et perspectives

Le but de ce stage était l'analyse de trajectoires en virage, leur classification en trois classes proposées dans un livrable du projet DIVAS et la construction d'indicateurs de risques.

Les données obtenues ayant été mesurées en juin 2009 par une version encore perfectible de l'observatoire de trajectoires, leur traitement a mis en évidence le problème de la synchronisation des capteurs. Ainsi, ne restaient après nos filtres que 38% des trajectoires présentes dans la base de données initiale, décrites uniquement au centre du virage par des variables qui, pour certaines, ont dû être recalculées (*écart*, *accélération*, *position*), supprimées (*cap, angle au volant*) ou construites (*rayon de courbure*, *accélération latérale*).

Après avoir effectué une première analyse descriptive des trajectoires ne révélant pas d'incohérence dans ce nouveau jeu de données, nous avons construit des indicateurs. Notre intention initiale était de calculer ces indicateurs pour chaque portion du virage, mais les trajectoires résistant aux filtres n'étaient pas assez longues. Nous avons néanmoins pu les calculer au centre du virage, puis nous les avons classés pour en sélectionner les plus pertinents. Ces derniers ont composé les axes d'une ACP destinés à révéler un classement de nos trajectoires. Cette première typologie a mis de côté les trajectoires les plus exotiques, celles qui ne suivent pas le virage et coupent le marquage central.

Nous avons ensuite cherché à construire un indicateur de risque, fondé sur la position sur la voie par rapport aux positions les plus observées. Ces positions étaient déterminées par rapport à l'abscisse curviligne du marquage central précédemment interpolé. Un arbre de décision et une régression logistique ont mis en évidence le fait que ce risque pouvait difficilement être expliqué par les indicateurs liés à la vitesse. En revanche toutes les trajectoires exotiques de notre troisième classe se sont avérées risquées.

D'autres méthodes d'analyse et de classification de trajectoires découvertes lors d'une conférence de l'institut national d'études démographiques sur l'étude de trajectoires au sens socio-démographique ont ensuite été testées. Comme ces méthodes ne s'appliquent pour l'heure qu'à une variable, nous avons choisi d'étudier ici l'*accélération latérale* qui traduit la relation entre position et vitesse. Ainsi, le package TraMineR analysant l'évolution de variables qualitatives, nous avons d'abord dû discrétiser notre variable, pour ensuite obtenir une classification en trois classes dont la dernière rappelait clairement les caractéristiques du troisième type de conduite proposé dans le livrable : une trajectoire "extérieur-intérieur-extérieur" pour éviter de trop décélérer.

Une classification très proche de celle-ci a été trouvée à l'aide du package R KML, qui applique, quant à lui, la méthode des K-moyennes à des données longitudinales quantitatives. Ces deux méthodes proposent des mises en forme des classes qui explicitent clairement leurs différences.

Enfin, une recherche sur les distances employées pour comparer des trajectoires nous a conduit à essayer la distance Dynamic Time Warping, distance permettant de prendre en compte la translation ou la compression de données longitudinales. Nous avons ainsi recouru à la classification ascendante hiérarchique par la méthode de Ward à partir d'une matrice des distances DTW, pour obtenir des classes encore assez proches de celles obtenues par les deux méthodes précédentes.

Nous attendions pour cette étude d'autres données que celles dont nous avons disposé, trop peu nombreuses et limitées au centre du virage. Notre travail fut donc de proposer des indicateurs et des méthodes dont la pertinence pourra être évaluée à partir de données plus conséquentes. En effet, nous ne connaissons pas à l'issue de cette analyse, parmi les méthodes de classification testées, la méthode qui s'adapte le mieux aux trajectoires en virage (classification ascendante hiérarchique avec la méthode de Ward, classification descendante hiérarchique avec la proc Varclus du logiciel SAS ou classification par la méthode des K-moyennes). En outre, il nous est difficile de déterminer l'intérêt de la distance Dynamic Time Warping, ou de savoir si notre indice de risque est significatif.

Des méthodes d'apprentissage statistique pourraient proposer, si les trajectoires avaient recouvert la totalité du virage, une prédiction de la trajectoire entière à partir de l'entrée dans le virage.

Enfin l'expérience de l'analyse des trajectoires prises par une vingtaine de conducteurs connus, pourrait élargir le champs de l'étude en intégrant des données psychologiques.

En conclusion, ce sont les méthodes plus que les résultats obtenues que nous avons rapportées. Il est envisageable qu'appliquées à de nouvelles trajectoires, elles conduisent à des résultats plus significatifs.

Bibliographie

Gabadinho A., Studer M., N M. et Ritschard G. (2010), *Traminer*. Rapport technique.

Genolini C. (2009), *K-means for longitudinal data (kml)*. Rapport technique.

Giorgino T. (2009), *Dynamic time warping algorithms*. Rapport technique.

Goyat Y., Auder F. et Menant F. (2008a), *Analyse des trajectoires pratiquées par les usagers sur les sites expérimentaux radarr*. Rapport technique.

Goyat Y., Auder F. et Menant F. (2008b), *Rapport de mise en oeuvre de l'observatoire de trajectoires sur différents sites - conclusion sur l'accidentologie des vl en courbe*. Rapport technique.

Goyat Y. et Menant F. (2008), *Concept de la configuration de la base de données associée à l'observatoire de trajectoires*. Rapport technique.

Goyat Y. (2008). *Estimation précise des trajectoires de véhicule par un système optique*. UNIVERSITÉ BLAISE PASCAL - CLERMONT II.

Milligan W. et Cooper M. (1985), *An examination of procedures for determining the number of clusters in a data set*. Rapport technique.

Nakache J.-P. (2004). *Approche pragmatique de la classification :*.

Saporta G. (2006). *Probabilités, analyse des données et statistique*.

Tenenhaus M. (2007). *Statistique : méthodes pour décrire, expliquer et prévoir*. Dunod.

Tufféry S. (2007). *Data mining et statistique décisionnelle*.